Biophysiological Mental-State Monitoring during Human-Computer Interaction

H A B I L I T A T I O N S S C H R I F T

zur Erlangung der Lehrbefähigung

für das Fach Informatik

vorgelegt dem Rat der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Humboldt-Universität zu Berlin

von
Dr. rer. nat. Thea Radüntz

Präsidentin der Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:
Prof. Dr. Elmar Kulke

Gutachter:
1. Prof. Dr. Beate Meffert
2. Prof. Dr. Hans-Martin Hasselhorn
3. Prof. Dr. Michael Falkenstein

Berlin, den 15.07.2020

So eine Arbeit wird eigentlich nie fertig,
man muss sie für fertig erklären,
wenn man nach Zeit und Umständen
das Mögliche getan hat.

Johann Wolfgang von Goethe
Italienische Reise, 2. Teil, 1787.

Abstract

The long-term negative consequences of inappropriate mental workload on employee health constitute a serious problem for a digitalized society. Continuous, objective assessment of mental workload can provide an essential contribution to the identification of such improper load. Recent improvements in sensor technology and algorithmic methods for biosignal processing are the basis for the quantitative determination of mental workload.

Neuronal workload measurement has the advantage that workload registration is located directly there where human information processing takes place, namely the brain. Preliminary studies for the development of a method for neuronal workload registration by use of the electroencephalogram (EEG) have already been carried out [Rad16, Rad17]. For the field use of these findings, the mental workload assessment on the basis of the EEG must be evaluated and its reliability examined with respect to several conditions in realistic environments. A further essential requirement is that the method can be combined with the innovative technologies of gel free EEG registration and wireless signal transmission.

Hence, the presented papers include two investigations. Main subject of the first investigation are experimental studies on the usability of commercially-oriented EEG systems for mobile field use and system selection for the future work. Main subject of the second investigation is the evaluation of the continuous method for neuronal mental workload registration in the field. Thereby, a challenging application was used, namely the arrival management of aircraft. The simulation of the air traffic control environment allows the realisation of realistic conditions with different levels of task load. Furthermore, the work is well contextualized in a domain which is very sensible to human-factors research.

Contents

1 Motivation and Goal

The development of advanced information and communication technology as well as of highly interactive work environments and work assistance systems is unstoppable and aims at improving work conditions and human well-being. Nevertheless, employees complain about high mental workload and stress. Problems arise from information overload, frequent work interruptions, or from a multitude of irrelevant information [KK01, LCS03, NIO02]. Furthermore, repetitive and monotonous tasks as a result from automation and supervisory control may be accompanied by complacency, fatigue, reduced vigilance, and increased error rates [PMS93, PMM94, HR84, DGR03, MB08]. In addition to this, overload and underload are a safety risk for further persons [Str01]. Consequently, it is important to develop work environments that fit humans' cognitive abilities by optimizing workload conditions. Human-computer interaction is the research field concerned with such questions. Human-computer interaction was defined by the Association for Computing Machinery as "a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them" ([HBC+92], p. 5). The main aim of researchers in this field is the development of novel technologies that facilitate human-computer interaction. The mental-state information can be used by HCI researchers to control devices or to give feedback to the user to prevent undesired situations and enhance wanted effects. To realize the exchange of information between a human and a computer a brain-computer interface (BCI) is possible. BCIs can be categorized as active, reactive, or passive [ZK11]. If the brain activity is intentionally generated by the user, the BCI is called active. A reactive BCI is one where the brain activity is generated by predetermined external stimuli that causes a particular brain-signal reaction. In case of a brain activity not intentionally evoked by the user or experimenter, we speak about a passive BCI. Active and reactive BCIs are typically used to control applications whereas passive BCIs are often used for mental-state monitoring during arbitrary tasks. The increasing number of publications related to BCIs indicates an ever-growing interest in HCI systems where encoded brain activity from the user is used as an alternative channel to send information to a computer. Thereby, the construct of user's mental state comprises the level of arousal, fatigue, or mental workload [MTD+08]. In this work, we focus on the latter. Hence, for identifying workload peaks in the course of work, it is important to continuously assess the mental state over time.

Evaluation of mental workload by means of questionnaires is a widely used procedure. The main advantages are that registration effort is minimal and user acceptance is high. In order to get a quasi-continuous measurement, questionnaires can be assessed several times. However, a timely increased request decreases user acceptance and is not always possible. Problematic is also that the measurement can cause a change of mental work-

load, e.g. during a monotonous task where the person gets activated by answering the questionnaire.

Analysis of physiological signals offers the possibility of continuous mental-state monitoring. Frequently used bio-physiological workload indicators base on the activity of the brain, the cardiovascular system, the visual system, or the electro-dermal activity. Their main disadvantage is the acceptance of the technology employed for signal registration. These technologies can cause qualitative and quantitative impairments and reduce subject's compliance. Advancements regarding wireless transmission technology, however, have broadened the application of bio-signal registration. Mobile registration allows the investigation of new research questions as well as the development of a number of new mobile applications in areas like gaming, fitness, and sports. In this context the acceptance of the used measurement technology plays in important role and has to be known prior to its application during human-computer interaction. Today, users can identify and quantify their mental state directly there where human information processing takes place. Registration of the electroencephalogram (EEG) offers such a possibility. However, registration of the EEG was often coupled with issues that complicate its field use during human-computer interaction. These issues result first of all from signal's susceptibility to artifacts. For ensuring minimal artifacts, registration of the EEG traditionally is conducted in shielded labs using gel-based electrodes and wired connections between electrodes and amplifier. Currently, wireless and gel-free devices for mobile EEG registration appeared on the market. Consequently, the question arises if the emerging devices are able to cope with the requirements of mobile technology. First and foremost, these requirements include the assurance of an appropriate signal quality of the registered EEG. As different applications have different quality needs, the development of a methodology for signal-quality comparison is particularly important. Such a methodology does not exist, although new devices continually appear on the market and must be benchmarked accordingly. Additionally, further important aspects regarding mobile use of EEG devices are user's requirements regarding wearing comfort, practicability, and design. Chapter 2 of the present work deals with both issues mentioned above.

Many years of research regarding analysis of the EEG and results from numerous studies as well as own preparatory work provided a solid guidance for the development of a method for continuous mental workload registration. Exemplary articles in that field are [Kli99, GSL+98, GS00, Pfu97, SM95, MPSG01, Wil01, Wil02, BLL+07, KDB+07, BHE+12]. It has been shown that variations in the power spectra showed the highest explanatory power. The reported results indicated that increased task load leads to an increase of the θ-frequency band power at frontal electrodes and a decrease of the α-frequency band power at parietal electrodes. This was confirmed by own studies and constituted the background for the development of a new method [Rad17] which was developed under laboratory conditions and provided good results. In particular, it could be proven that continuous registration and quantification of mental workload in the classes low, moderate, and high was possible during execution of cognitive tasks. As a next step, the new method had to be evaluated under field conditions. This general research question is subject of the chapter 3. The method evaluation took place

under simulated work conditions at the DLR in Braunschweig. The main aim of the experimental research at the DLR was to demonstrate that mental workload registration in a sensitive work area can lead to important insights. An optimal workload level implies beneficial effects for the performance and health of the employees whereas a suboptimal one has the opposite effects. The selected air traffic controller working place provided adequate conditions for such kind of research. The simulator environment allowed for a systematic variation of task load. During arrival management procedures, we were able to vary the traffic volume in a broader range and to generate exceptional events. Additionally, we were able to investigate the planning abilities and working memory capacity of air traffic controllers.

To sum up, the long-term goal is the use of the new method for objective registration of mental workload during human-computer interaction. For this, two research problems had to be solved: the selection of a registration system acceptable for mobile field use and the evaluation and application of the new workload-registration and indexing method under realistic work conditions outside the lab.

2 Evaluation of Commercially-Oriented EEG Systems for Mobile Field Use

2.1 Related Work and Research Objectives

In the previous years, registration of brain activity by means of the EEG has become more and more popular not only in science but also in the home and gaming sector. Mobile EEG amplifiers, wireless signal transmission, and dry-sensor technology provide the opportunity to transfer BCI applications from the laboratory to daily-life environments [HYW+17, VKED14]. Moreover, mobile EEG registration enables a number of new applications related to mental-state monitoring and brain-behavior relationship, e.g. in sports [WMK19].

Nowadays, progress in sensor technology enables the production of low-cost, lightweighted, marketable, and – above all – mobile devices. However, various problems still hamper an extended use of the emerging EEG systems. In the following, two research questions are addressed: signal quality and wearing comfort. Both problems have not been satisfactory solved yet and there have only been a few articles dealing with them.

Good signal quality and repeatability of the measurements are main requirements for the interpretation of the registered brain activity. A common way to achieve this is the registration of the EEG in a shielded lab, preparation of the subject's skin before the electrodes are placed to reduce the impedance, and the use of wired signal transmission. Unfortunately, these standard procedures limit the usability of an EEG device and narrow its application outside the lab.

Over the last few years, research engineers and EEG system manufacturers have been working on overcoming these issues and allowing easy and reliable EEG registration outside the lab. Emerging sensor technology enables gel-free EEG registration and thus quick and easy application of EEG devices by the users themselves. The main problem of such dry electrodes is the assurance of a good signal quality. Using gel electrodes, a low impedance between electrodes and skin is easily realized because of their permanent contact to the skin because of the properties of the gel. This becomes particularly difficult to achieve for dry electrodes that work without the conductive gel. The transition to wireless transmission of large amounts of data is less problematic. Signal transmission from several channels and with high resolution are not an issue any more.

Regarding signal quality, there were studies that focused primarily on the evaluation of mobile vs. non-mobile devices, which neglected the emerging dry-electrode systems [FAD+13, RTV+14]. Other investigations concentrated only on a single dry-electrode device and considered its general performance [RJA+16, CDT15]. Another study dealt with one dry-electrode and one gel-based device [DCP+13]. However, the

majority of the articles described self-developed dry sensors and compared their signal quality to that of a traditional gel-based system (e.g., [NKC10, GKAE12]). An interesting study that examined more than two devices included a wireless gel-based device, wireless saline-based device, wired dry-electrode device, and wired gel-electrode device [GLL+15].

Nonetheless, even the best signal quality is worthless, if the users are not willing to wear the devices over a longer period of time because of usability issues. Only a small number of studies were concerned with devices' wearing comfort and design requirements (e.g., [IOS+16, NLG+15]). The studies indicated that for assuring user acceptance, devices should be lightweight, comfortable, not painful to wear, and with an unobtrusive design. Limitations of these studies were a limited number of participants, lack of comparisons among different devices, or a too short wearing duration of the EEG headsets. Most of the studies focused primarily on wearing comfort and neglected user-experience aspects such as emotional design.

To recap, there is growing interest among users in brain-activity monitoring and increased efforts for developing mobile EEG devices. Signal quality and user acceptance of new devices are crucial and their evaluation constitutes an important subject of research. Radüntz (2018) [Rad18] contributes to the evaluation of signal quality and provides a methodical approach for the comparison of different devices. Radüntz et al. (2019) [RM19**?**] deal with user experience and provide a systematic procedure for its assessment. The obtained results build an important prerequisite for the general application of the emerging devices outside the lab, enable functionality assessments of new devices, and support their further development.

2.2 Experimental Design

The investigation was conducted with seven mobile and mainly gel-free EEG devices in a non-shielded office setting. The 24 subjects participating completed over the course of 9 consecutive workdays a total of 9 sessions. The first session was aimed at familiarizing the subjects with the experiment. In the following 7 days, one EEG device per day was selected in random order and tested independently of the others. The subjects wore the device for 60 min and performed the same sequence of tasks. They played computer games, performed one easy and one more demanding cognitive task for 5 min each, and 1-min rest measurements with eyes closed and eyes opened.

2.3 Signal Quality Evaluation

To evaluate the signal quality, it was analyzed in the time and in the frequency domain. In the time domain the proportion of artifacts and signal-to-noise ratio (SNR) of the devices were examined [Rad18]. For determining the proportion of artifacts, a visual examination of the signals and a statistical analysis were performed. The visual examination is a widely applied and well-accepted method and was conducted by a medical technical assistant with specialization in EEG analysis. To statistically evaluate the

proportion of artifacts, we conducted an analysis of variance (ANOVA) with a repeated measures design. As a second criterion, the SNR was used as a reliable and often used instrument for quality evaluation of bio-signals. For statistical evaluation of the SNR values a non-parametric Friedman test of the differences among the six devices was calculated.

The gel-based g.LADYbird device yielded the best results for both criteria, followed by the g.SAHARA device that had the best performance among the gel-free devices. Remarkably, for none of the gel-free devices the signal was greater than the noise, i.e. the SNR values were less than 0 dB. This could prove to be particularly problematic if precise measurements are required.

While the evaluation in the time domain aimed at the first instance to identify the very obvious differences regarding the devices' artifact susceptibility, the evaluation in the frequency domain went a step further. After removing all of the artifact-contaminated segments, a deeper look at the signal was necessary to determine whether it reflected the actual brain activity. If the devices recorded a brain signal with appropriate quality, the analysis of the signal's frequency band power is a common method for hypothesis evaluation. For the signal quality evaluation in the frequency domain three generally known facts were examined: the Berger effect, the increase in the frontal θ-frequency band power, and the decrease in the parietal α-frequency band power as task demands became greater. Statistical evaluation of these three facts was conducted using Wilcoxon paired difference tests for each EEG system.

All three phenomena were proven to be true for the gel-based g.LADYbird device. The g.SAHARA and BR8+ devices were able to capture significant differences regarding the Berger effect and the decrease in the parietal α-band power during the demanding task. The Jellyfish device was the only one among the gel-free devices that was able to register a significant increase in the θ-band power during the demanding task. Only two devices did not measure any significant changes in signal's band power: the MindCap and the Trilobite devices.

2.4 User Experience Evaluation

For user experience evaluation, subjects were asked how long they would be able to wear the EEG headset at the end of each session. Furthermore, they answered questions regarding the device's design and rated its practicability. During the last session, paired comparisons were conducted between every 21 pairs of two devices presented. Participants were asked to select the headset that they were willing to wear over a longer period of time or even daily. Finally, subjects completed a questionnaire where they had to rank the devices regarding wearing comfort and visual appearance separately.

To assess the user experience, questionnaires were developed with items regarding the wearing comfort and the visual appearance. The answers provided insight into the relation between user experience aspects and device preference [RM19, RR18]. The wearing comfort given by a device was the main factor for its daily use. The visual appearance of the device was certainly a further important point. It became influential

when comfort was assured. Users were not willing to accept less comfort for a more attractive headset design. Subject's mood and headset pressure were measured by the psycho-physiological method of cross-modality matching [**?**]. They were related to each other and changed over the wearing time. This alternation was particularly prominent for the Trilobite device where the changes in the course of time became significant. In contrast, the g.LADYbird device seemed to be the most comfortable. Results also indicated that head pressure was a mediator between device properties and subject's mood, with device's weight as significant predictor.

2.5 Discussion

The evaluation of EEG systems for mobile field use was done under two aspects. First, the signal quality has to meet the requirements of the particular application. Such applications allow for human-computer interaction and are established not only in the research sector but also in BCI, the home and gaming sector, and sports. Second, the wearing comfort and the visual appearance have to meet the needs of the user. The conducted experiments and their analyses allowed conclusions about both.

The proposed procedure for signal quality evaluation used signal properties from the time domain (artifact proportion, SNR) as well as frequency domain (band power). The combination of the parameters from both domains proved to be well suited for determining the different signal quality of the EEG devices. As expected, outstanding performance was obtained for the traditional gel-based but mobile device. None of the other emerging devices could reach its signal quality. The remaining devices did not meet all requirement of an appropriate signal quality, although some developers could decide to use them for mobile applications where precise measurements are not required. Furthermore, study's results indicated that gel-free EEG devices manufactured only in one size can lead to bad outcomes regarding the signal quality.

The methodical approach for user experience evaluation used questionnaires and cross-modality matching. It was appropriate for assessing comfort, practicability, and design aspects. Based on it, we were able to draw important conclusions for the further development of mobile EEG systems. Developers should be aware of potential comfort issues that could arise in the course of time because of the pin electrodes and should attach importance to the weight of the headset for assuring comfort and well-being. They should also consider possible interaction effects between the weight, electrode type, and the number of electrodes. To provide practical information to users of EEG devices, the signal quality and user experience results must be combined for concluding which system could be used under which condition.

To sum up, the developed framework proved to be suitable for the comparison of different EEG devices. The rapid development of the EEG-equipment market requires easily applicable benchmarking methods. Only in this way, it becomes possible to choose the most proper EEG system for a particular application. Future results of new EEG systems could be easily compared with the findings of our study that provide a procedure for the evaluation of further emerging EEG technology. The current study widened the

state of the art by providing metrics for signal quality and user experience evaluation.

3 Mental-State Monitoring of Air Traffic Controllers

3.1 Related Work and Research Objectives

One central topic of mental-state monitoring research is the assessment of mental workload. Mental workload describes the cognitive demands required in order to solve a task and relates them to the cognitive resources available [Kah73]. Following this definition, it can be expected that registration and evaluation of mental workload is particular important in order to minimize errors during the work to be done, optimize human performance, and enhance mental health. In general, high mental workload may arise from the inability to cope with increasing task load [EWKD91, Wic02] but also from a simultaneous interaction of task load with emotional aspects [ACD+04] and individual's training and experience level [XS00].

High cognitive demands combined with high responsibility during work are linked to high mental workload. In such cases, employees have to maintain their performance even under difficult situations. Air traffic control is a typical example of such safety-critical environment. Here, inappropriate workload can have a number of negative consequences not only on employee's health but also on the safety of persons. This is why assurance of an optimal workload range is highly important. Investigating if the employees are operating within their personal optimal cognitive-performance range is therefore especially necessary. The design, implementation, and evaluation of interactive computing systems for human use depend on the findings from this research. Thus, a valid and reliable method for registering mental workload is urgently needed.

Methods for registering mental workload are categorized into subjective and objective methods. The subjective measurements use traditional questionnaires in order to assess subject's experienced workload. The objective determination of mental workload is based on behavioural data and physiological parameters.

Questionnaires' main advantages were the simplicity of assessment and high user acceptance. However, subjective measurements were problematic because of their susceptibility to subjective distortion, social-desirability restrictions regarding the appropriateness of the answer, and subject's inability to introspect. Furthermore, it is important that the registration method does not interact with the task or alter subject's mental state by imposing additional demands as it is the case during subjective assessment of workload by means of questionnaires. The subjective measurements are usually categorized in one-dimensional and multi-dimensional questionnaires. One-dimensional questionnaires consist of only one rating scaling and provide a general workload rating. They are easy to understand by the subject and can be quickly conducted during the task sev-

eral times. In contrast, multi-dimensional questionnaires include more than one rating scale, their assessment takes longer, and does not allow a frequent repeatability during the task. Much more, they are conducted at the end of the task. The final-rating value is computed by aggregation of the single ratings of the scales.

Measurement of individual's performance on a task was another way to determine workload. Hereby, identification of workload rely on the relationship concept between workload and performance and implied that individual performance decreases under high mental workload. Studies also indicated that motivation, training, and experience could contribute to maintain performance at the same level by investing more effort and in this way mitigated the impact of workload [SMW+13, Mat01]. This means, that although performance during an easy and high-demanding task might be the same, the amount of experienced workload between both is different considering the cognitive resources needed for task solving. Thus, the increased mental workload cannot always be measured directly by means of performance break-down. Additionally, the workload should not only be detectable in retrospect or after the occurrence of errors as it is the case when performance measures are used for workload detection.

Recording and analysis of physiological signals offered insight into subject's psychophysiological state. The main idea underlying the assessment of workload using biosignals considered arousal and activation mechanisms of the organism reacting to the task load. Over the years, the various physiological parameters have been evaluated for their validity regarding continuous mental workload registration. Among them are the brain activity, cardiovascular parameters as well as occular data.

In a review article, Borghini et al. [BAV+14] provided a detailed overview of the measurement and analysis of neurophysiological signals for the determination of mental workload and confirmed essentially the relations known to date. In general, the most common brain activity registration techniques are the electroencephalography, functional near infrared spectroscopy (fNIRS), functional magnetic resonance imaging (fMRI), and magnetoencephalography (MEG). Each of them comprises advantages and disadvantages linked to their operating principle. An advantage of fMRI and MEG is the fast preparation time whereas EEG and NIRS require application of electrode caps. A further advantage of fMRI and MEG is a spatial resolution of millimeters that is not limited to cortical areas. The spatial resolution of the EEG and fNIRS depends on the number of sensors used, can be measured in centimeters, and is limited below cortical surface. One drawback with fMRI is its temporal resolution. It takes several seconds for the blood flow to change. Similarly, the hemodynamic activity detected by fNIRS is a delayed representation of the cortical activity. In the contrary, the temporal resolution of EEG and MEG is in the range of milliseconds. A major advantage of the EEG and fNIRS are their mobility while fMRI and MEG are fixed at the structure of the building. The fNIRS is the technique which is less sensitive to movement artifacts, followed by the EEG where innovative artifact-rejection algorithms contribute to a better signal quality. The MEG and fMRI are highly sensitive to movements. Furthermore, the EEG causes low costs, fNIRS moderate, fMRI high, and MEG the highest. Taken together, for mobile, affordable, and accessible insights about brain function and mental state with a high temporal resolution, the EEG is the method of choice.

Basically, changes in the α-frequency and θ-frequency band powers of the EEG related to mental workload have been confirmed many times and proved to be meaningful in accordance with the findings of the last 50 years. These EEG bands were linked to different levels of workload (e.g. [BHE+12, CSP+12, LR11]) and showed a decrease of the α-frequency band power and an increase of the θ-frequency band power with increasing mental workload.

The majority of workload studies was conducted in laboratories and dealt with the analysis of the EEG during cognitive tasks related to working memory and executive control (e.g., [Kli99, GSL+98]). Some authors investigated whether a brain-state monitoring was possible on the basis of universal and general activation signs in the EEG (e.g., [BBY14, KQH+14], while others tested the possibilities and limitations of machine learning algorithms. The research problem of interest was the possible transferability of the classifier over tasks or individuals. The answer could be found by cross-task or cross-subject training, respectively (e.g., [BP12, PB12]). According to Kohlmorgen et al. [KDB+07], a universally applicable workload detector with fixed parameters did not seem to be realistic at the moment. The selection of appropriate data for classifier's training needs more elucidation. Borghini et al. [BAV+14] further concluded that no convincing algorithms were available for a reliable online workload detection. In order to avoid overfitting and increase the stability of the classifier performance over time a smaller number of features would be beneficial [ABF+15].

In a previous study, a mental-workload classifier that does not need retraining, neither for new subjects nor for new tasks, was developed [Rad17]. The so-called Dual Frequency Head Maps (DFHM) were developed in a laboratory study during execution of well-established cognitive tasks. The head maps consist of personalized spectral features and their spatial occurrence (i.e., frontal θ-band and parietal α-band powers). Support vector machines are used for classification in three classes: low, moderate, or high workload. Under laboratory conditions, it was successfully proved that the DFHM method is universally applicable with fixed parameters for mental-workload indexing.

For its practical application at the workplace, it is also important that its applicability is examined not only in the laboratory but also under more realistic conditions. For this, a study in cooperation with the German Aerospace Center was conducted focusing on air traffic controllers. The main aim of this study was the evaluation of reliability and reproducibility of the DFHM method's results under realistic conditions.

3.2 Experimental Design

The selected environment for the investigation was the radar controller approach position that was simulated at the Air Traffic Management and Operations Simulator of the German Aerospace Center. The air traffic management simulation environment reproduced a regular air traffic controller working place with all typical components such as a radar screen, a weather display, and a voice communication system to talk to pilots [MRT+18]. The simulator experiment was expected to be representative for real operations because of its similarity to controllers' working environment and controllers'

communication with pseudo pilots. The sample consisted of 21 subjects between the ages of 22 and 64 years that were not only DLR internal research controllers but also controllers from the German and Austrian air navigation service provider.

The traffic intensity was varied in the range between 25 and 55 aircraft per hour (ac/h) related to four simulation scenarios. To induce additional workload, we implemented four additional simulation scenarios where one aircraft requested priority due to a sick passenger on board after the 10th minute of the simulation. The eight scenarios (four traffic intensities with and four without priority request flight) had a duration between 20 min and 25 min. The recorded data during the simulations included the EEG, heart rate, instantaneous self-assessment (ISA) and NASA-TLX questionnaires, as well as air traffic performance and radio-communication data.

3.3 Subjectively Experienced Workload

Analysis of the subjectively experienced workload aimed to find out if it was the number of aircraft or the occurrence of an exceptional event that stressed controllers the most and if there was an interaction effect between both. While the effect of the number of aircraft was evident by the questionnaire methods, the impact of the priority-flight event remained doubtful [RFT+19]. Controllers' ISA and NASA-TLX ratings showed only a weekly significant discrimination between sessions with and without priority flight using standard ANOVA tests.

3.4 Indexing Mental Workload Using Dual Frequency Head Maps

The DFHM index showed highly significant correlations between scenarios with similar traffic-load conditions [RFMM20]. The DFHM index was also able to assess significant differences between the different levels of air traffic volume, with exception of the neighboring levels of 35 and 45 ac/h. Considered over all subjects, the DFHM index did not reveal significant differences regarding the priority-flight request.

More insight regarding intra-individual differences linked to the DFHM-workload index was gained from subject clustering by means of the subjectively experienced workload differences during the scenarios [FRM20]. In this way, we were able to obtain highly significant interaction effects between subjective workload-cluster affiliation and traffic load as well as priority-flight request. For subjects reporting that they experienced workload variation between the different scenarios, the DFHM-workload index yielded significant differences between traffic-load levels and priority-flight request conditions. In contrast to the significant differences obtained for the workload-sensitive cluster, the DFHM-workload index behaved differently for the not-sensitive cluster and did not yield any significant differences for any of the factors.

Analyses of performance data emphasized these findings. Results revealed a tendency to more loss of separation and lower prioritization during the extreme traffic load

condition for workload-sensitive subjects that was less pronounced for the not-sensitive subjects.

3.5 The Effect of Planning on Mental Workload

The relationship between the task demands and the experienced workload is strongly mediated by the operator's individual characteristics [Hil04]. Air traffic control is a complex task with high cognitive demands including three main cognitive processes: planning, evaluation, and monitoring [PBCL96]. Past research identified working-memory requirements as the most important component of the mental workload arising during air traffic control [ACLV16]. However, planning abilities will be the dominating and most crucial cognitive skill for meeting the demands of the growing air traffic volume required in the future [DKD00]. Thus, planning effort will mainly contribute to the experienced mental workload.

In the study by Radüntz [Rad20], the DFHM method was applied for capturing mental workload objectively during a planning and a working memory task. As a planning task the Tower of Hanoi (TOH) was employed, as a working memory task the automated orientation span (AOSPAN) task.

The DFHM-workload index was significantly higher for the TOH than for AOSPAN task suggesting that more cognitive resources were required during planning than working memory task. The result was consistent with literature that stated that planning is a higher-order executive function that integrates core cognitive processes such as working memory, inhibitory control, and cognitive flexibility [ÁPB+15, Dia13, MFE+00]. During the learning phase of the TOH task, we were able to obtain a significant interaction effect between task load and working memory capacity on mental workload. Thereby, mental workload of subjects with higher working memory capacity significantly decreased while the workload of subjects with lower working memory capacity did not yield significant changes. The effect was particularly prominent for the mental workload assessed by the EEG whereas the number of errors and planning time showed only a weak tendency in that direction.

Furthermore, the issue of unclear goal states during planning was of particular interest. In an additional investigation of the TOH with suboptimal sequences of sub-goals, we gained proof that tasks with a higher level of goal ambiguity induce higher mental workload [RFM20].

3.6 Discussion

The submitted articles addressed questions of mental-state monitoring in field use. In particular, the functionality of the recently developed DFHM method was evaluated regarding stability of results and generalization properties of the DFHM-workload index. Of special interest were method's inter-individual and cross-task abilities. In conclusion, it can be stated that a reliable determination of mental workload in a realistic setting and with real-world scenarios was possible. The DFHM-workload index from the EEG

was able to differentiate between the various load conditions that arose from the traffic volumes. However, it was only limitedly able to differentiate between load conditions arising from the priority request or from the interaction between traffic volume and priority request. This observation fitted well to controllers' reports. Most of them mentioned that during the scenarios with low and medium traffic volumes they had no difficulties to deal with the priority request. During the scenarios with higher traffic demands the situation changed and the priority-flight event became more demanding. Nevertheless, some controllers were able to easily handle the situation. The objectively measured workload assessed by the DFHM method corresponded to controller's subjectively experienced workload. Performance data provided an additional indicator that subjects from the workload-sensitive cluster experienced more workload compared to the others. In future work more research is needed in order to understand which individual factors contribute to these interpersonal differences in the perception of workload.

In this context, a topic of particular interest is related to problem solving and planning and the identification of factors decisive for human-computer interaction. Understanding the interrelation among them may contribute to adjust conditions, facilitate learning, enhance planning, and reduce workload in accordance to the cognitive abilities of the individual. The results of the TOH experiment indicated an initial learning process on neurological level that might produce behavioral changes after longer practice. This fits well with the assumption by Hardy and Wright [HW18] that mental workload reflects the cognitive abilities of the performer, captures individual differences, and reveals additional information about the cognitive state although task performance might be similar. Additionally, the issue of goal ambiguity is of particular interest for problem solving in digitized working environments that often comprise suboptimal sequences of sub-goals because of an unclear goal state. Based on our results, application developers can be advised to predefine and communicate the sequence of the sub-goals to the users using intelligent assistance systems for reducing ambiguity.

A limitation of our study was the realization of the exceptional event as recurring priority-flight request. The surprising effect of the unexpected event might have diminished after the first occurrence of the request. Thereafter, air traffic controllers might have adjusted their strategy and behavior in order to be prepared to appropriately react to a recurring event. Studies that aim to understand the effect of an unexpected event on workload, should pay more attention on this issue. Finally, a larger sample size would be beneficial.

4 Conclusions

With the development and availability of low-cost and easy-to-use systems for the registration and analysis of bio-physiological signals, a monitoring of the physiological and mental state of a person is possible. Based on this, the range of application of bio-signal analysis was significantly extended, in particular regarding HCI. In this context, mental workload monitoring is of special interest for assuring an appropriate workload in accordance to user's cognitive capacity. Prevention of over or under load contributes to an optimal performance and better health. In order to adjust conditions, knowledge about user's mental state is required. Thus, mental workload registration is necessary. The present work deals with this issue. Subject of research was an application of the EEG for indexing mental workload.

The numerous studies published after the year 2000 were fairly different, depending on the specific question, purpose, and expertise of the authors. Nevertheless, the authors agreed that the EEG was a suitable instrument for identifying the current level of mental workload. Analysis in the frequency domain is especially relevant because certain frequency bands of the EEG have proven to be particularly informative and were therefore being used more and more frequently for mental-workload detection. The developed method using DFHM makes use of these well-established parameters for quantifying mental workload by calculating a workload index. Its great advantage is that it is valid for different subjects and tasks.

The submitted articles addressed questions related to the DFHM-workload index and its applicability and functionality outside the laboratory. The first study provided scientific feedback regarding the claims of manufacturers of emerging EEG technology. The results are of particular interest for researchers that want to use the new wearable devices for their studies. It paves the way for the development of commercial applications of wearables and contributes to progress in consumer health informatics and health-enabling technologies. In addition, the results provided a guidance for the technological development direction of new EEG devices for applications in human-computer interactions and the development of brain-computer interfaces.

The second study was about the validation of an EEG-based mental-workload index, determined by the new DFHM method. Previously developed and tested in laboratory settings, it was now transferred to a field use. Thereby, a challenging application was chosen, namely the arrival management of aircraft. The air traffic control environment allowed the realisation of realistic conditions with different levels of task load. Furthermore, the work was well contextualized in a domain, the air traffic management, which is very sensible to human-factors research. The results showed that the EEG-based workload index was sensitive to the cognitive load as related to the different experimental tasks and was also confirming the workload scoring self-assessed of the subjects through

the online questionnaire.

The developed framework for bio-physiological mental-state monitoring enables further fundamental research linked to different aspects of human-computer interaction. The capability of the DFHM index from the EEG to successfully register mental workload suggests it as a useful tool for further studies. In future research, we aim at employing the DFHM index for human-computer interaction research related to mental-workload issues of the modern society. Although the temporal resolution of the EEG permits a determination in the range of seconds, the states to be detected originate from long-running procedures and therefore require further research about an informative time frame for averaging classifier's output. Future promising applications of the DFHM-workload index include research about effects of human-computer interaction, human factors, ergonomic designs linked to the mental state for development and testing new interfaces, determination of the effectiveness of training and simulation programs, or even the characterization of group dynamics by collecting synchronous EEG data from multiple subjects. A real-time application of EEG parameters to determine vigilance, emotion, workload, and stress as examples of mental states is closely related to a number of new and interesting research questions. These can help to improve the living and working conditions of the single individual but also of the human community.

5 Submitted Articles

1. **Thea RADÜNTZ (2018)**: Signal Quality Evaluation of Emerging EEG Devices. Frontiers in Physiology, 9:98, doi: 10.3389/fphys.2018.00098

2. **Thea RADÜNTZ**, Beate MEFFERT **(2020)**: Cross-modality Matching for Evaluating User Experience of Emerging Mobile EEG Technology. IEEE Transactions on Human-Machine Systems – accepted

3. **Thea RADÜNTZ**, Beate MEFFERT **(2019)**: User Experience of 7 Mobile Electroencephalography Devices: Comparative Study. JMIR Mhealth Uhealth 2019; 7(9):e14474, DOI: 10.2196/14474

4. **Thea RADÜNTZ**, Norbert FÜRSTENAU, André TEWS, Lea RABE, Beate MEFFERT **(2019)**: The Effect of an Exceptional Event on the Subjectively Experienced Workload of Air Traffic Controllers. In: Longo L., Leva M. C. (eds) 'Human Mental Workload: Models and Applications', H-WORKLOAD 2018, Revised Selected Papers. Communications in Computer and Information Science, vol. 1012. Springer International Publishing, eBook ISBN 978-3-030-14273-5, DOI 10.1007/978-3-030-14273-5_14, pp. 239-257

5. **Thea RADÜNTZ**, Norbert FÜRSTENAU, Thorsten MÜHLHAUSEN, Beate MEFFERT **(2020)**: Indexing Mental Workload during Simulated Air Traffic Control Tasks by Means of Dual Frequency Head Maps. Frontiers in Physiology, 11:300, doi: 10.3389/fphys.2020.00300

6. **Thea RADÜNTZ (2020)**: The Effect of Planning, Strategy Learning, and Working Memory Capacity on Mental Workload. Nature Scientific Reports 10, 7096, doi: 10.1038/s41598-020-63897-6

Bibliography

[ABF+15] Arico, Pietro ; Borghini, Gianluca ; Flumeri, Gianluca D. ; Colosimo, Alfredo ; Graziani, Ilenia ; Imbert, Jean-Paul ; Granger, Geraud ; Benhacene, Railene ; Terenzi, Michela ; Pozzi, Simone ; Babiloni, Fabio: Reliability over time of EEG-based mental workload evaluation during Air Traffic Management (ATM) tasks. In: *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, aug 2015

[ACD+04] Averty, Philippe ; Collet, Christian ; Dittmar, André ; Athènes, Sylvie ; Vernet-Maury, Evelyne: Mental workload in air traffic control: an index constructed from field tests. In: *Aviation, space, and environmental medicine* 75 (2004), Nr. 4, S. 333–341

[ACLV16] Arbula, Sandra ; Capizzi, Mariagrazia ; Lombardo, Nicoletta ; Vallesi, Antonino: How Life Experience Shapes Cognitive Control Strategies: The Case of Air Traffic Control Training. In: *PLOS ONE* 11 (2016), 06, Nr. 6, 1-17. http://dx.doi.org/10.1371/journal.pone.0157731. – DOI 10.1371/journal.pone.0157731

[ÁPB+15] Ávila, Rafaela T. ; Paula, Jonas J. ; Bicalho, Maria A. ; Moraes, Edgar N. ; Nicolato, Rodrigo ; Malloy-Diniz, Leandro F. ; Diniz, Breno S.: Working Memory and Cognitive Flexibility Mediates Visuoconstructional Abilities in Older Adults with Heterogeneous Cognitive Ability. In: *Journal of the International Neuropsychological Society* 21 (2015), may, Nr. 5, S. 392–398. http://dx.doi.org/10.1017/s135561771500034x. – DOI 10.1017/s135561771500034x

[BAV+14] Borghini, Gianluca ; Astolfi, Laura ; Vecchiato, Giovanni ; Mattia, Donatella ; Babiloni, Fabio: Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. In: *Neuroscience & Biobehavioral Reviews* 44 (2014), 58-75. http://dx.doi.org/10.1016/j.neubiorev.2012.10.003. – DOI 10.1016/j.neubiorev.2012.10.003. – ISSN 0149–7634. – Applied Neuroscience: Models, methods, theories, reviews. A Society of Applied Neuroscience (SAN) special issue.

[BBY14] Bashivan, Pouya ; Bidelman, Gavin M. ; Yeasin, Mohammed: Spectrotemporal dynamics of the EEG during working memory encoding and maintenance predicts individual behavioral capacity. In: *European Journal*

of Neuroscience 40 (2014), oct, Nr. 12, S. 3774–3784. http://dx.doi.org/10.1111/ejn.12749. – DOI 10.1111/ejn.12749

[BHE+12] Brouwer, A. M. ; Hogervorst, M. A. ; Erp, J. B. F. ; Heffelaar, T. ; Zimmerman, P. H. ; Oostenveld, R.: Estimating workload using EEG spectral power and ERPs in the n-back task. In: *Journal of Neural Engineering* 9 (2012), jul, Nr. 4, 045008. http://dx.doi.org/10.1088/1741-2560/9/4/045008. – DOI 10.1088/1741–2560/9/4/045008

[BLL+07] Berka, C. ; Levendowski, D. J. ; Lumicao, M. N. ; Yau, A. ; Davis, G. ; Zivkovic, V. T. ; Olmstead, R. E. ; Tremoulet, P. D. ; Craven, P. L.: EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks. In: *Aviat Space Environmental Medicine* 78(5, Suppl.) (2007), S. B231–B244

[BP12] Baldwin, Carryl L. ; Penaranda, B. N.: Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification. In: *NeuroImage* 59 (2012), Nr. 1, 48-56. http://dx.doi.org/10.1016/j.neuroimage.2011.07.047. – DOI 10.1016/j.neuroimage.2011.07.047. – ISSN 1053–8119. – Neuroergonomics: The human brain in action and at work

[CDT15] Callan, D. E. ; Durantin, G. ; Terzibas, C.: Classification of single-trial auditory events using dry-wireless EEG during real and motion simulated flight. In: *Frontiers in Systems Neuroscience* 9 (2015), 11. http://dx.doi.org/10.3389/fnsys.2015.00011. – DOI 10.3389/fnsys.2015.00011. – ISSN 1662–5137

[CSP+12] Capilla, Almudena ; Schoffelen, Jan-Mathijs ; Paterson, Gavin ; Thut, Gregor ; Gross, Joachim: Dissociated α-band modulations in the dorsal and ventral visual pathways in visuospatial attention and perception. In: *Cerebral Cortex* 24 (2012), Nr. 2, S. 550–561

[DCP+13] Duvinage, M. ; Castermans, T. ; Petieau, M. ; Hoellinger, T. ; Cheron, G. ; Dutoit, T.: Performance of the Emotiv Epoc headset for P300-based applications. In: *BioMedical Engineering OnLine* 12 (2013), Jun, Nr. 1, 56. http://dx.doi.org/10.1186/1475-925X-12-56. – DOI 10.1186/1475–925X–12–56. – ISSN 1475–925X

[DGR03] Debitz, U. ; Gruber, H. ; Richter, G.: *Psychische Gesundheit am Arbeitsplatz. Teil 2: Erkennen, Beurteilen und Verhüten von Fehlbeanspruchungen.* Bd. 2. 3. InfoMediaVerlag, 2003

[Dia13] Diamond, Adele: Executive Functions. In: *Annual Review of Psychology* 64 (2013), jan, Nr. 1, S. 135–168. http://dx.doi.org/10.1146/annurev-psych-113011-143750. – DOI 10.1146/annurev–psych–113011–143750

[DKD00] Dittmann, Andrea ; Kallus, K. W. ; Damme, Dominique V. A. N.: *Integrated Task and Job Analysis of Air Traffic Controllers - Phase 3: Baseline Reference of Air Traffic Controller Tasks and Cognitive Processes in the ECAC Area*. 2000

[EWKD91] *Kapitel* Workload assessment in multi-task environments. In: Eggemeier, F. ; Wilson, G. F. ; Kramer, A. F. ; Damos, D. L.: *Multiple-task performance*. Taylor & Francis, 1991, S. 207–216

[FAD+13] Forney, E. ; Anderson, C. ; Davies, P. ; Gavin, W. ; Taylor, B. ; Roll, M.: A Comparison of EEG Systems for Use in P300 Spellers by Users With Motor Impairments in Real-World Environments. In: *Proceedings of the Fifth International Brain-Computer Interface Meeting*, Graz University of Technology Publishing House, 2013

[FRM20] Fürstenau, N. ; Radüntz, T. ; Mühlhausen, T.: Model-based Development of a Mental Workload-sensitivity Index for Subject Clustering. In: *Theoretical Issues in Ergonomics Science* (2020). http://dx.doi.org/10.1080/1463922X.2020.1711990. – DOI 10.1080/1463922X.2020.1711990. – ISSN 1463–922X

[GKAE12] Guger, C. ; Krausz, G. ; Allison, B. ; Edlinger, G.: Comparison of Dry and Gel Based Electrodes for P300 Brain-Computer Interfaces. In: *Frontiers in Neuroscience* 6 (2012), 60. http://dx.doi.org/10.3389/fnins.2012.00060. – DOI 10.3389/fnins.2012.00060. – ISSN 1662–453X

[GLL+15] Grummett, T.S. ; Leibbrandt, R.E. ; Lewis, T.W. ; DeLosAngeles, D. ; Powers, D.M.W. ; Willoughby, J.O. ; Pope, K.J. ; Fitzgibbon, S.P.: Measurement of neural signals from inexpensive, wireless and dry EEG systems. In: *Physiological Measurement* 36 (2015), Nr. 7, 1469. http://stacks.iop.org/0967-3334/36/i=7/a=1469

[GS00] Gevins, A. ; Smith, M. E.: Neurophysiological Measures of Working Memory and Individual Differences in Cognitive Ability and Cognitive Style. In: *Cerebral Cortex* 10 (2000), Nr. 9, 829-839. http://dx.doi.org/10.1093/cercor/10.9.829. – DOI 10.1093/cercor/10.9.829. – Last accessed on 2014-02-18

[GSL+98] Gevins, A. ; Smith, M. E. ; Leong, H. ; McEvoy, L. ; Whitfield, S. ; Du, R. ; Rush, G.: Monitoring Working Memory Load during Computer-Based Tasks with EEG Pattern Recognition Methods. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 40 (1998), Nr. 1, 79-91. http://dx.doi.org/10.1518/001872098779480578. – DOI 10.1518/001872098779480578

[HBC+92] Hewett, Thomas T. ; Baecker, Ronald ; Card, Stuart ; Carey, Tom ; Gasen, Jean ; Mantei, Marilyn ; Perlman, Gary ; Strong, Gary ;

Verplank, William: ACM SIGCHI Curricula for Human-Computer Interaction. New York, NY, USA : Association for Computing Machinery, 1992. – Forschungsbericht. – ISBN 0897914740

[Hil04] Hilburn, Brian: *Cognitive complexity in air traffic control: a literature review*. 01 2004

[HR84] Hacker, W. ; Richter, P.: *Psychische Fehlbeanspruchung. Psychische Ermüdung, Monotonie, Sättigung und Stress (Spezielle Arbeits- und Ingenieurpsychologie in Einzeldarstellungen)*. Bd. 2. 2. Berlin: Springer, 1984

[HW18] Hardy, David J. ; Wright, Matthew J.: Assessing workload in neuropsychology: An illustration with the Tower of Hanoi test. In: *Journal of Clinical and Experimental Neuropsychology* 40 (2018), may, Nr. 10, S. 1022–1029. http://dx.doi.org/10.1080/13803395.2018.1473343. – DOI 10.1080/13803395.2018.1473343

[HYW+17] Huang, X. ; Yin, E. ; Wang, Y. ; Saab, R. ; Gao, X.: A mobile EEG system for practical applications. In: *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2017. – ISSN null, S. 995–999

[IOS+16] Izdebski, Krzysztof ; Oliveira, Anderson S. ; Schlink, Bryan R. ; Legkov, Petr ; Kärcher, Silke ; Hairston, W. D. ; Ferris, Daniel P. ; König, Peter: Usability of EEG Systems: User Experience Study. In: *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. New York, NY, USA : ACM, 2016 (PETRA '16). – ISBN 978–1–4503–4337–4, 34:1–34:4

[Kah73] Kahneman, D.: *Attention and Effort*. Prentice-Hall, Englewood Cliffs, 1973

[KDB+07] Kohlmorgen, J. ; Dornhege, G. ; Braun, M. L. ; Blankertz, B. ; Müller, K. R. ; Curio, G. ; Hagemann, K. ; Bruns, A. ; Schrauf, M. ; Kincses, W. E.: Improving Human Performance in a Real Operating Environment through Real-Time Mental Workload Detection. In: Dornhege, G. (Hrsg.) ; R. Millán, J. del (Hrsg.) ; Hinterberger, T. (Hrsg.) ; McFarland, D. (Hrsg.) ; Müller, K.R. (Hrsg.): *Towards Brain-Computer Interfacing*, MIT Press, Cambridge, 2007, S. 409–422

[KK01] Kompier, M. A. J. ; Kristensen, T. S.: Organisational Work Stress Interventions in a Theoretical, Methodological and Practical Context. In: Dunham, J. (Hrsg.): *Stress in the Workplace: Past, Present and Future*. London : Whurr Publishers, 2001, S. 164–190

[Kli99] Klimesch, W.: EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. In: *Brain Research Reviews* 29 (1999), Nr. 2-3, S. 169–195

[KQH+14] Ke, Yufeng ; Qi, Hongzhi ; He, Feng ; Liu, Shuang ; Zhao, Xin ; Zhou, Peng ; Zhang, Lixin ; Ming, Dong: An EEG-based mental workload estimator trained on working memory task can work well under simulated multi-attribute task. In: *Frontiers in Human Neuroscience* 8 (2014), 703. http://dx.doi.org/10.3389/fnhum.2014.00703. – DOI 10.3389/fnhum.2014.00703. – ISSN 1662–5161

[LCS03] Landsbergis, P. A. ; Cahill, J. ; Schnall, P.: The Changing Organisation of Work and the Safety and Health of Working People: A Commentary. In: *Journal of Occupational Environmental Medicine* 45 (2003), Nr. 1, S. 61–72. http://dx.doi.org/10.1097/00043764-200301000-00014. – DOI 10.1097/00043764–200301000–00014

[LR11] Lei, S. ; Roetting, M.: Influence of Task Combination on EEG Spectrum Modulation for Driver Workload Estimation. In: *Human Factors* 53(2) (2011), S. 168–179

[Mat01] Matthews, Gerald: Levels of transaction: a cognitive sciences framework for operator stress. In: Hancock, P. A. (Hrsg.) ; Desmond, P. A. (Hrsg.): *Stress, Workload, and Fatigue*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, 2001, S. 5–33

[MB08] May, J. F. ; Baldwin, C. L.: Driver Fatigue: The Importance of Identifying Causal Factors of Fatigue when Considering Detection and Countermeasure Technologies. In: *Transportation Research Part F* 12 (2008), S. 218–224. http://dx.doi.org/10.1016/j.trf.2008.11.005. – DOI 10.1016/j.trf.2008.11.005. – Last accessed on 2011-11-03

[MFE+00] Miyake, A. ; Friedman, N. P. ; Emerson, M. J. ; Witzki, A. H. ; Howerter, A. ; Wager, T. D.: The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. In: *Cognitive Psychology* 41 (2000), aug, Nr. 1, 49 - 100. http://dx.doi.org/http://dx.doi.org/10.1006/cogp.1999.0734. – DOI http://dx.doi.org/10.1006/cogp.1999.0734. – ISSN 0010–0285. – Last accessed on 2014-03-17

[MPSG01] McEvoy, Linda K. ; Pellouchoud, Emiliana ; Smith, Michael E. ; Gevins, Alan: Neurophysiological signals of working memory in normal aging. In: *Cognitive Brain Research* 11 (2001), Nr. 3, 363-376. http://dx.doi.org/10.1016/S0926-6410(01)00009-X. – DOI 10.1016/S0926–6410(01)00009–X. – ISSN 0926–6410

[MRT+18] Mühlhausen, Thorsten ; Radüntz, Thea ; Tews, André ; Gürlük, Hejar ; Fürstenau, Norbert: Research Design to Access the Mental Workload of Air Traffic Controllers. Version: oct 2018. http://dx.doi.org/10.1007/978-3-030-02053-8_63. In: *Human Systems Engineering and*

Design. Springer International Publishing, oct 2018. – DOI 10.1007/978–3–030–02053–8_63, S. 415–421

[MTD+08] Müller, Klaus-Robert ; Tangermann, Michael ; Dornhege, Guido ; Krauledat, Matthias ; Curio, Gabriel ; Blankertz, Benjamin: Machine learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring. In: *Journal of Neuroscience Methods* 167 (2008), Nr. 1, 82-90. http://dx.doi.org/https://doi.org/10.1016/j.jneumeth.2007.09.022. – DOI https://doi.org/10.1016/j.jneumeth.2007.09.022. – ISSN 0165–0270. – Brain-Computer Interfaces (BCIs)

[NIO02] NIOSH, NORA Organization of Work Team Members: The Changing Organization of Work and the Safety and Health of Working People / National Institute for Occupational Safety and Health (NIOSH). 2002 (2002-116). – Forschungsbericht

[NKC10] Nikulin, V. V. ; Kegeles, J. ; Curio, G.: Miniaturized Electroencephalographic Scalp Electrode for Optimal Wearing Comfort. In: *Clinical Neurophysiology* 121 (2010), S. 1007–1014. http://dx.doi.org/10.1016/j.clinph.2010.02.008. – DOI 10.1016/j.clinph.2010.02.008

[NLG+15] Nijboer, Femke ; Laar, Bram van d. ; Gerritsen, Steven ; Nijholt, Anton ; Poel, Mannes: Usability of Three Electroencephalogram Headsets for Brain-Computer Interfaces: A Within Subject Comparison. In: *Interacting with Computers* 27 (2015), 07, Nr. 5, 500-511. http://dx.doi.org/10.1093/iwc/iwv023. – DOI 10.1093/iwc/iwv023. – ISSN 0953–5438

[PB12] Penaranda, B. N. ; Baldwin, Carryl L.: Temporal Factors of EEG and Artificial Neural Network Classifiers of Mental Workload. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 56 (2012), Nr. 1, 188-192. http://dx.doi.org/10.1177/1071181312561016. – DOI 10.1177/1071181312561016

[PBCL96] Pawlak, William S. ; Brinton, Christopher R. ; Crouch, Kimberly ; Lancaster, Kenneth M.: A Framework For The Evaluation Of Air Traffic Control Complexity, 1996

[Pfu97] Pfurtscheller, Gert: EEG event-related desynchronization (ERD) and synchronization (ERS). In: *Electroencephalography and Clinical Neurophysiology* 103 (1997), Nr. 1, S. 26

[PMM94] Parasuraman, R. ; Mouloua, M. ; Molloy, R.: Monitoring Automation Failures in Human Machine Systems. In: Mouloua, M. (Hrsg.) ; Parasuraman, R. (Hrsg.): *Human Performance in Automated Systems: Curent Research and Trends*. Hillsdale, NJ: Earlbaum, 1994, S. 45–49

[PMS93] Parasuraman, R. ; Molloy, R. ; Singh, I. L.: Performance Consequences of Automation Induced Complacency. In: *International Journal of Aviation Psychology* 3 (1993), S. 1–23

[Rad16] Radüntz, T.: *Kontinuierliche Bewertung psychischer Beanspruchung an informationsintensiven Arbeitsplätzen auf Basis des Elektroenzephalogramms*. Berlin, Germany, Humboldt-Universität zu Berlin, Department of Computer Science, Ph. D. thesis, 2016. http://edoc.hu-berlin.de/docviews/abstract.php?id=42402

[Rad17] Radüntz, Thea: Dual Frequency Head Maps: A New Method for Indexing Mental Workload Continuously during Execution of Cognitive Tasks. In: *Frontiers in Physiology* 8 (2017), 1019. http://dx.doi.org/10.3389/fphys.2017.01019. – DOI 10.3389/fphys.2017.01019. – ISSN 1664–042X

[Rad18] Radüntz, Thea: Signal Quality Evaluation of Emerging EEG Devices. In: *Frontiers in Physiology* 9 (2018), 98. http://dx.doi.org/10.3389/fphys.2018.00098. – DOI 10.3389/fphys.2018.00098. – ISSN 1664–042X

[Rad20] Radüntz, Thea: The Effect of Planning, Strategy Learning, and Working Memory Capacity on Mental Workload. In: *Nature Scientific Reports* 10 (2020), apr, Nr. 1. http://dx.doi.org/10.1038/s41598-020-63897-6. – DOI 10.1038/s41598–020–63897–6

[RFM20] Radüntz, Thea ; Freyer, Marion ; Meffert, Beate: Ambiguous Goals During Human-Computer Interaction Induce Higher Mental Workload. In: Harris, Don (Hrsg.) ; Li, Wen-Chin (Hrsg.): *Engineering Psychology and Cognitive Ergonomics. Mental Workload, Human Physiology, and Human Energy*. Cham : Springer International Publishing, 2020. – ISBN 978–3–030–49044–7, S. 81–90

[RFMM20] Radüntz, Thea ; Fürstenau, Norbert ; Mühlhausen, Thorsten ; Meffert, Beate: Indexing Mental Workload during Simulated Air Traffic Control Tasks by Means of Dual Frequency Head Maps. In: *Frontiers in Physiology* 11 (2020), apr. http://dx.doi.org/10.3389/fphys.2020.00300. – DOI 10.3389/fphys.2020.00300

[RFT+19] Radüntz, Thea ; Fürstenau, Norbert ; Tews, André ; Rabe, Lea ; Meffert, Beate: The Effect of an Exceptional Event on the Subjectively Experienced Workload of Air Traffic Controllers. Version: Januar 2019. http://dx.doi.org/10.1007/978-3-030-14273-5_14. In: *Human Mental Workload: Models and Applications*. Springer, Januar 2019. – DOI 10.1007/978–3–030–14273–5_14. – ISBN 978–3–030–14272–8

[RJA+16] Rogers, J.M. ; Johnstone, S.J. ; Aminov, A. ; Donnelly, J. ; Wilson, P. H.: Test-retest reliability of a single-channel, wireless EEG system. In: *International Journal of Psychophysiology* 106 (2016), Nr. Supplement

C, 87 - 96. http://dx.doi.org/https://doi.org/10.1016/j.ijpsycho.2016.06.006. – DOI https://doi.org/10.1016/j.ijpsycho.2016.06.006. – ISSN 0167–8760

[RM19] Radüntz, Thea ; Meffert, Beate: User Experience of 7 Mobile Electroencephalography Devices: Comparative Study. In: *JMIR Mhealth Uhealth* 7 (2019), Sep, Nr. 9, e14474. http://dx.doi.org/10.2196/14474. – DOI 10.2196/14474. – ISSN 2291–5222

[RR18] Radüntz, Thea ; Rose, Uwe: Influence of Personal Characteristics and Device Properties on Wearable's Rank Order. In: Karwowski, Waldemar (Hrsg.) ; Ahram, Tareq (Hrsg.): *Intelligent Human Systems Integration*. Cham : Springer International Publishing, 2018. – ISBN 978–3–319–73888–8, S. 321–326

[RTV+14] Ries, A.J. ; Touryan, J. ; Vettel, J. ; McDowell, K. ; Hairston, W.: A Comparison of Electroencephalography Signals Acquired from Conventional and Mobile Systems. In: *Journal of Neuroscience and Neuroengineering* 3 (2014), Nr. 1, S. 10–20. http://dx.doi.org/doi:10.1166/jnsne.2014.1092. – DOI doi:10.1166/jnsne.2014.1092

[SM95] Sterman, M. B. ; Mann, C. A.: Concepts and applications of EEG analysis in aviation performance evaluation. In: *Biological Psychology* 40 (1995), Nr. 1, 115-130. http://dx.doi.org/10.1016/0301-0511(95)05101-5. – DOI 10.1016/0301–0511(95)05101–5. – ISSN 0301–0511. – EEG in Basic and Applied Settings

[SMW+13] Saxby, D. J. ; Matthews, G. ; Warm, J. S. ; Hitchcock, E. M. ; Neubauer, C.: Active and Passive Fatigue in Simulated Driving: Discriminating Styles of Workload Regulation and Their Safety Impacts. In: *Journal of Experimental Psychology: Applied* 19 (2013), Nr. 4, S. 287–300. http://dx.doi.org/http://doi.org/10.1037/a0034386. – DOI http://doi.org/10.1037/a0034386

[Str01] Sträter, O.: Warum passieren menschliche Fehler und was kann man dagegen tun? In: *Forum PrÃ¤vention* AUVA - Allgemeine Unfallversicherungsanstalt, Wien, 2001

[VKED14] Vos, Maarten de ; Kroesen, Markus ; Emkes, Reiner ; Debener, Stefan: P300 speller BCI with a mobile EEG system: Comparison to a traditional amplifier. In: *Journal of neural engineering* 11 (2014), 04, S. 036008. http://dx.doi.org/10.1088/1741-2560/11/3/036008. – DOI 10.1088/1741–2560/11/3/036008

[Wic02] Wickens, C. D.: Multiple resources and performance prediction. In: *Theoretical Issues in Ergonomics Science* 3 (2002), Nr.

2, 159-177. http://dx.doi.org/10.1080/14639220210123806. – DOI 10.1080/14639220210123806

[Wil01] Wilson, G. F.: In-flight Psychophysiological Monitoring. In: Fahrenberg, F. (Hrsg.) ; Myrtek, M. (Hrsg.): *Progress in Ambulatory Monitoring*. Hogrefe and Huber, 2001, S. 435–454

[Wil02] Wilson, G. F.: Psychophysiological Test Methods and Procedures. In: Charlton, S.G. (Hrsg.) ; O'Brien, T.G. (Hrsg.): *Handbook of Human Factors Testing and Evaluation*, Lawrence Erlbaum Associates, 2002, S. 127–156

[WMK19] Wang, Chun-Hao ; Moreau, David ; Kao, Shih-Chun: From the Lab to the Field: Potential Applications of Dry EEG Systems to Understand the Brain-Behavior Relationship in Sports. In: *Frontiers in Neuroscience* 13 (2019), 893. http://dx.doi.org/10.3389/fnins.2019.00893. – DOI 10.3389/fnins.2019.00893. – ISSN 1662–453X

[XS00] Xie, B. ; Salvendy, G.: Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments. In: *Work & Stress* 14(1) (2000), S. 74–99

[ZK11] Zander, T. O. ; Kothe, C.: Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general. In: *Journal of Neural Engineering* 8 (2011), Nr. 2, 025005. http://stacks.iop.org/1741-2552/8/i=2/a=025005

Statements of Authorship and Originality

I hereby certify that the thesis I am submitting is entirely my own original work except where otherwise indicated. I am aware of the University's regulations concerning plagiarism, including those regulations concerning disciplinary actions that may result from plagiarism. Any use of the works of any other author, in any form, is properly acknowledged at their point of use.

Berlin, 13th of May 2020 Thea Radüntz

Danksagung

Mein Dank gilt zunächst der Bundesanstalt für Arbeitsschutz und Arbeitsmedizin, die mir die Möglichkeit eröffnete, im Rahmen meiner Arbeitstätigkeit in der Gruppe 3.4 'Mentale Gesundheit und kognitive Leistungsfähigkeit' unter der Leitung von Frau Dr. Gabriele Freude, zu promovieren und zu habilitieren. Die wissenschaftlichen Vorarbeiten der Fachgruppe bilden die Grundlage dieser Arbeit. Besonderer Dank gilt darüber hinaus allen Kollegen, die bei der Konzeption und Durchführung der experimentellen Untersuchungen aktiv mitgearbeitet haben.

Die letzten vier Jahre hatte ich die Gelegenheit, mit wunderbaren Menschen in Berlin und Braunschweig zusammenzukommen, die ein offenes Ohr für fachliche Diskussionen hatten, bei organisatorischen Problemen behilflich waren, die viele hilfreiche Impulse gaben sowie meine Ideen kritisch hinterfragten. Ihnen allen möchte ich herzlichst danken.

Frau Prof. Dr. Beate Meffert danke ich von ganzem Herzen. Seit Beginn meines Studiums der Informatik, über die Betreuung meiner Diplomarbeit, die Betreuung meiner Dissertation bis zur Abgabe meiner Habilitation erhielt ich von ihr jede erdenkliche, hilfreiche Unterstützung und durfte viele anregende Diskussionen mit ihr führen. Ihre Hilfe kam mir über die vergangenen 20 Jahre in zahlreichen Angelegenheiten sehr zugute. Sie ist die beste Mentorin, die man sich wünschen kann.

frontiers in Physiology

TECHNOLOGY REPORT
published: 14 February 2018
doi: 10.3389/fphys.2018.00098

Signal Quality Evaluation of Emerging EEG Devices

*Thea Radüntz**

Mental Health and Cognitive Capacity, Work and Health, Federal Institute for Occupational Safety and Health, Berlin, Germany

Electroencephalogram (EEG) registration as a direct measure of brain activity has unique potentials. It is one of the most reliable and predicative indicators when studying human cognition, evaluating a subject's health condition, or monitoring their mental state. Unfortunately, standard signal acquisition procedures limit the usability of EEG devices and narrow their application outside the lab. Emerging sensor technology allows gel-free EEG registration and wireless signal transmission. Thus, it enables quick and easy application of EEG devices by users themselves. Although a main requirement for the interpretation of an EEG is good signal quality, there is a lack of research on this topic in relation to new devices. In our work, we compared the signal quality of six very different EEG devices. On six consecutive days, 24 subjects wore each device for 60 min and completed tasks and games on the computer. The registered signals were evaluated in the time and frequency domains. In the time domain, we examined the percentage of artifact-contaminated EEG segments and the signal-to-noise ratios. In the frequency domain, we focused on the band power variation in relation to task demands. The results indicated that the signal quality of a mobile, gel-based EEG system could not be surpassed by that of a gel-free system. However, some of the mobile dry-electrode devices offered signals that were almost comparable and were very promising. This study provided a differentiated view of the signal quality of emerging mobile and gel-free EEG recording technology and allowed an assessment of the functionality of the new devices. Hence, it provided a crucial prerequisite for their general application, while simultaneously supporting their further development.

Keywords: signal quality, electroencephalogram (EEG), mobile EEG, dry electrodes, wearables

OPEN ACCESS

Edited by:
Kris Thielemans,
University College London,
United Kingdom

Reviewed by:
Tommaso Gili,
Enrico Fermi Center, Italy
Danilo Mandic,
Imperial College London,
United Kingdom

***Correspondence:**
Thea Radüntz
raduentz.thea@baua.bund.de

Specialty section:
This article was submitted to Biomedical Physics, a section of the journal Frontiers in Physiology

Received: *11 October 2017*
Accepted: *29 January 2018*
Published: *14 February 2018*

Citation:
Radüntz T (2018) Signal Quality Evaluation of Emerging EEG Devices. Front. Physiol. 9:98. doi: 10.3389/fphys.2018.00098

1. INTRODUCTION

Electroencephalogram (EEG) registration as a direct measurement of brain activity has unique potentials. The fact that all physical and mental processes are controlled by our brain suggests that such information is also reflected in the registered signal. Hence, an EEG is one of the most reliable and predicative indicators when studying human cognition, evaluating a subject's health condition, or monitoring their mental state.

A main requirement for the interpretation of the registered brain activity is good signal quality. A common way to achieve this is the registration of the EEG in a shielded lab and preparation of the subject's skin before the electrodes are placed to reduce the impedance. Unfortunately, these standard procedures limit the usability of an EEG device and narrow its application outside the lab. An additional challenge when it comes to real-life applications involves the wired connections

from the electrode cap to an amplifier and computer. These severely restrict a subject's mobility and decrease user acceptance of the measuring technique.

Over the last few years, research engineers and EEG system manufacturers have been working on overcoming these issues and allowing easy and reliable EEG registration outside the lab. By means of wireless signal transmission, they have developed mobile devices that allow subjects to move more freely. Furthermore, emerging sensor technology allows gel-free EEG registration and enables quick and easy application of EEG devices by the users themselves. However, the signal quality of these new devices remains unclear.

There have only been a few articles dealing with this issue. Among these, there were studies that focused primarily on the evaluation of mobile vs. non-mobile devices, which neglected the emerging dry-electrode systems (Forney et al., 2013; Ries et al., 2014). Other investigations concentrated only on a single dry-electrode device and considered its general performance (Callan et al., 2015; Rogers et al., 2016). Finally, there were studies on one dry-electrode and one gel-based device (Zander et al., 2011; Johnstone et al., 2012; Duvinage et al., 2013). However, the majority of the articles described self-developed dry sensors and compared their signal quality to that of a traditional gel-based system (Sullivan et al., 2007; Nikulin et al., 2010; Grozea et al., 2011; Saab et al., 2011; Debener et al., 2012; Guger et al., 2012). An interesting study that examined more than two devices included a wireless gel-based device, wireless saline-based device, wired dry-electrode device, and wired gel-electrode device (Grummett et al., 2015). To the best of our knowledge, no signal comparison studies of several wireless dry-electrode systems are available.

In our work, we compared the signal quality of various mobile and gel-free EEG devices. Hence, our study offers a differentiated look at nascent EEG recording technology and enables functionality assessments of the new devices. The obtained results build a crucial prerequisite for the general application of the emerging devices outside the lab and simultaneously support their further development.

2. MATERIALS AND EXPERIMENTS

2.1. EEG Systems

The investigation focused on six mobile EEG devices. They are illustrated in **Figure 1**, and their specifications are summarized in **Table 1**.

The EPOC is the only device in our study that works with saline-based, wet felt sensors. It has two reference electrodes that are mounted at the parietal sides (P3/P4 locations).

The Jellyfish is also an easy-to-apply device. It consists of a headband with four dry electrodes and an adhesive reference electrode at the mastoid. The four electrodes can be applied at either frontal or parietal sites. The manufacturer recommends the use of foam-based electrodes for the frontal sites and spring-loaded electrodes for the parietal sites (**Figure 1E**). In our study, we registered the frontal EEG and thus attached foam-based electrodes to the headband.

The Trilobite device comes from the same manufacturer as the Jellyfish. It includes three foam-based frontal electrodes and 29 spring-loaded pin electrodes. Additionally, the device has a ground electrode and reference ear-clip electrode.

The BR8+ device comprises two frontal foam-based electrodes and six spring-loaded pin electrodes. Ground and reference electrodes are applied with ear-clips. The ear pads of the device do not have any technical functionality.

The pin electrodes of g.tec's g.SAHARA/g.Nautilus device are mounted on a traditional EEG cap. Adhesive ground and reference electrodes are applied at the mastoids. The cap of the device comes in small, medium, and large sizes. We only employed the medium-size cap in order to reduce the financial cost.

Finally, we also included a traditional, gel-based but mobile EEG system, the g.LADYbird/g.Nautilus device by g.tec. It includes 16 active electrodes and an ear-clip electrode as a reference. Although the cap size can vary, just as with the g.SAHARA/g.Nautilus device, we only used the medium-size cap in our study to reduce the cost. The g.LADYbird/g.Nautilus device was primarily developed for research and medical use. We included it to our study as a state-of-the-art reference for EEG registration in relation to the signal quality.

It was not possible to use the same sample rate for every device. In order to maintain comparable conditions for the later evaluation, we attempted to operate the devices with sample rates that were as similar as possible. Hence, for the Jellyfish and Trilobite devices, the EEG was registered at 256 Hz, and the g.SAHARA and g.LADYbird devices used 250 Hz. For both of the remaining devices, manual adjustment of the sample rate to 250 Hz was not possible. Thus, we had to run the EPOC device at 128 Hz and the BR8+ device at 1000 Hz. Furthermore, we applied a digital notch filter at 50 Hz during all of the recordings. All of the EEG devices utilized wireless signal transmission to a computer.

2.2. Procedure and Subjects

Our study was conducted in a non-shielded office setting. Twenty-four subjects (11 females and 13 males, 26–66 years of age, with a mean age of 42.8) participated in the study. They tested one device per day for 60 min. During this time, the participants played computer games and performed one easy and one more demanding cognitive task for 5 min each. The 0-back task represented the easy task, where subjects were instructed to press the mouse button if the letter "X" appeared on the screen (Kirchner, 1958; Gazzaniga et al., 2013). The stop signal task was a more demanding inhibition task (Logan, 1994; Dimoska, 2005). During this task, the subjects were instructed to press the green mouse button as fast as possible if a horizontal left arrow was presented on the screen and the red mouse button if a horizontal right arrow appeared. If a horizontal arrow was quickly followed by a vertical arrow, they were instructed to inhibit their response and not press either button. They had to respond as quickly as possible and remember that their main aim was to keep the frame around the arrow green. A red frame meant that they were too slow. Hence, if it was red, they had to speed up their response while still paying attention to the vertical arrow.

Finally, we conducted two rest measurements, where we instructed the subjects to sit quietly for a minute, first with their

FIGURE 1 | Six mobile EEG devices tested in our study.

eyes open and subsequently with their eyes closed. The devices were selected in random order over the participants and days, while the sequence of the performed tasks remained constant for all.

All of the investigations conducted were approved by the local review board of our institution, and the experiments were conducted in accordance with the Declaration of Helsinki. All of the procedures were carried out with the adequate understanding and written consent of the subjects.

3. METHODS

To evaluate the signal quality, we examined the proportion of artifacts and signal-to-noise ratio of the devices in the time domain and considered the signal properties in the frequency domain.

3.1. Evaluation in Time Domain

Two hypotheses were postulated based on our expectations for the signal quality in regard to the time domain. In order to test both hypotheses, we employed EEG data from all of the computer tasks.

3.1.1. Proportion of Artifacts

Hypothesis 1: The gel-based device has a significantly lower proportion of artifacts than the gel-free devices.

TABLE 1 | Technical data of tested EEG devices (n.s.: not specified).

Device	EPOC	Trilobite	Jellyfish	BR8+	g.SAHARA	g.LADYbird
Electrode type	Wet (saline)	Dry (spring, foam)	Dry (spring, foam)	Dry (spring, foam)	Dry (pins)	Gel (active)
No. of channels	14	32	4	8	16	16
Battery life [hours]	12	10	10	11	10	10
Resolution [bit]	14	24	24	24	24	24
Max. sample rate [Hz]	128	500	500	1000	500	500
Bandwidth [Hz]	0.2–45	0.23–n.s.	0.23–n.s.	0.12–125	0.1–40	0.1–40
Weight [g]	116	524	95	269	233	165

The evaluation of the EEG in the time domain with regard to hypothesis 1 was conducted manually. The visual inspection and discarding of contaminated EEG segments by an expert is a widely applied and well-accepted method in research and clinical settings. Therefore, we asked for assistance from a medical technical assistant (MTA) with specialization in EEG analysis and years of experience in that field.

The MTA visually inspected the EEGs of each subject from all of the devices and manually marked artifact segments using a skill-based state-of-the-art procedure. Thereby, she did not mark physiological artifacts (e.g., eye blinks, eye movements) because these were not related to the device properties.

We then computed the percentage of denoted artifacts compared to the entire recording time for each channel. Finally, we calculated the means over the channels and subjects for each device.

3.1.2. Signal-to-Noise Ratio

Hypothesis 2: The gel-based device has a significantly higher signal-to-noise ratio than the gel-free devices.

We computed the signal-to-noise ratio (SNR) as a standard method to assess the signal quality. The SNR values were calculated using the following relation:

$$SNR = 10 \cdot \log_{10}\left(\frac{\sigma_x^2}{\sigma_e^2}\right) \text{ [dB]} \quad (1)$$

where σ_x^2 is the variance of the signal, and σ_e^2 is the variance of the noise. For zero mean signals, as found here, this results in the following:

$$SNR = 10 \cdot \log_{10} \frac{\sum_{(i=1)}^{N} x_i^2}{\sum_{(i=1)}^{N} (s_i - x_i)^2} \quad (2)$$

where N is the number of sample points, x_i is the noise reduced signal at time i, and s_i is the band-pass filtered signal at time i.

First, we filtered the original raw signals using a Hamming band-pass filter (order 100) between 1 and 40 Hz and obtained the filtered signal s_i. Subsequently, we applied the artifact subspace reconstruction (ASR) algorithm to calculate the noise-reduced signal amplitudes x_i (Mullen et al., 2013). This algorithm is particularly suitable for cleaning continuous, non-triggered data from artifacts. Furthermore, the approach is well established within the scientific community (e.g., Bulea et al., 2015; Luu et al., 2017) and recommended for wireless, dry-electrode systems (Mullen et al., 2015). In the following, we give a brief description of how the algorithm works.

The algorithm identifies a clean signal segment from the given EEG and computes its statistics. Next, the ASR runs with a sliding window over the EEG and conducts a principal component analysis for each window. It removes high-variance components with three standard deviations above the mean and reconstructs their content using a mixing matrix calculated from the previously identified clean segment. For a more detailed explanation of the mathematical background and functionality of the algorithm, we advise the interested reader to consult the appropriate articles by the developers.

For the residual noise signal in the denominator, we used the difference between band-pass filtered signal s_i and the noise-reduced signal from the ASR algorithm, x_i. The signal quality of the devices could be compared under this assumption. For each device, the SNR values were computed for all of the electrodes and subjects.

3.2. Evaluation in Frequency Domain

To evaluate the signal quality in the frequency domain, we formulated three more hypotheses. We expected that if a device had good signal quality, we would be able to measure significant differences in the signal's frequency band power values for the various tasks.

Hypothesis 3: For devices with good signal quality, a significant Berger effect can be obtained between measurements with the eyes open and those with the eyes closed.

Our third hypothesis was based on the so-called Berger effect (Berger, 1929). This states that the parietal alpha band power is supposed to be smaller with the eyes open than closed. This is also known as the "alpha block."

For each device, we considered the two rest measurements with the eyes open and closed. We removed all of the segments previously marked as artifacts. We subsequently applied a Hamming band-pass filter for the alpha frequency band (8–12 Hz) to the artifact-free signals of the parietal electrodes (**Figure 2**). The relative band power values were averaged over the electrodes for the rest measurements with the eyes open and closed.

Hypothesis 4: For devices with good signal quality, a significant increase in the frontal theta power can be obtained when comparing the easy and more demanding cognitive tasks.

The fourth hypothesis was based on the dependency of the frontal theta band power on the experienced workload. Based on the results from numerous previous investigations (e.g., Gevins et al., 1998; Radüntz, 2016), we expected a significant increase in the frontal theta power when comparing the easy and more demanding cognitive tasks.

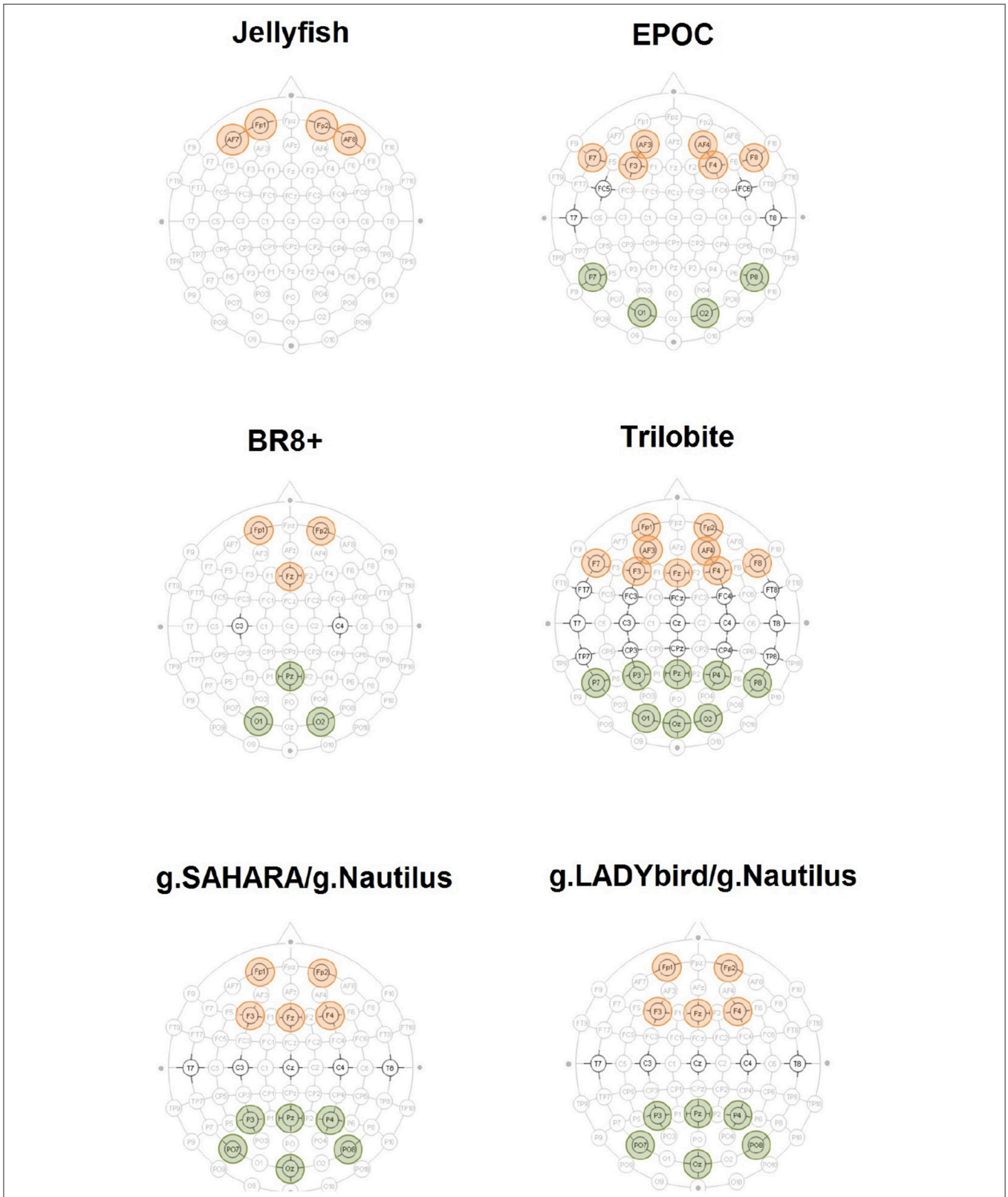


FIGURE 2 | Accentuated positions constitute EEG devices' layout. The aggregated electrodes for the frontal theta-band power evaluation are highlighted in red. The electrodes used for the parietal alpha-band power calculation are highlighted in green.

To this end, we focused on the EEGs from the 0-back and stop signal tasks of each device. First, we removed all of the previously marked artifact segments. We subsequently applied a Hamming band-pass filter for the theta frequency band (4–8 Hz) to the artifact-free signals of the frontal electrodes (**Figure 2**). The relative band power values were averaged over the electrodes for both the 0-back and stop signal tasks.

Hypothesis 5: For devices with good signal quality, a significant decrease in the parietal alpha band power can be obtained when comparing the easy and more demanding cognitive tasks.

Our last hypothesis was also based on findings regarding the experienced workload, but now with respect to the parietal alpha band power, which is expected to significantly decrease when comparing the easy and more demanding cognitive tasks (Gevins et al., 1998; Radüntz, 2016).

For each device, we considered the EEGs from the 0-back and stop signal tasks. We removed all of the previously marked artifact segments and applied a Hamming band-pass filter for the alpha frequency band (8–12 Hz) to the artifact-free signals of the parietal electrodes (**Figure 2**). Next, the relative band power values were averaged over the electrodes for both the 0-back and stop signal tasks.

4. RESULTS

Digital signal processing was performed with MATLAB. All of the statistical calculations were carried out using SPSS. Furthermore, we provide Supplementary Material with the subjects' values for each analysis and system.

4.1. Evaluation in Time Domain

4.1.1. Proportion of Artifacts

To statistically evaluate the proportion of artifacts for the various devices, we conducted an analysis of variance (ANOVA) with a repeated measures design. The six devices constituted the levels used for testing each subject at each level of the within-subject variable. Bonferroni's corrected *post-hoc* tests were conducted to determine the differences between the levels.

The results are presented in **Figure 3**. They indicate significant differences among the devices in relation to their proportions of artifact-contaminated signal segments [Greenhouse-Geisser: $F_{(2.72;\ 62.61)} = 15.88$, $p < 0.001$]. The *post-hoc* tests showed that the traditional gel-based g.LADYbird device had significantly fewer artifacts than almost all of the other devices, and that the BR8+ device had significantly more artifacts than most of the others. The dry pin-electrode device (g.SAHARA) yielded a significantly lower artifact proportion than the remaining pin-electrode devices. However, it had a higher proportion than the gel-based device. Finally, no significant differences compared to any other device could be obtained for the EPOC device.

4.1.2. Signal-to-Noise Ratio

Before going into detail about the SNR results, it should be noted that the ASR algorithm failed when examining the EEGs of four subjects that were recorded with the Trilobite device. This was because no segment of the needed length could be found as a reference for the algorithm, where all of the electrodes' signals were concurrently clean. Hence, these four subjects had to be excluded from the subsequent statistical computations for all the devices.

For each device, we calculated the median of the SNR values for each electrode over all the subjects and tasks. At the first site, we found obvious differences among the devices and noticed that g.LADYbird and g.SAHARA had the highest SNR values (**Figure 4**). In order to statistically evaluate these observations, we calculated the median of the SNR values over all the channels for each subject and device. We then conducted a non-parametric Friedman test of the differences among the six devices.

The results indicated significant differences in the devices' SNR values ($\chi^2 = 71.34$, $df = 5$, $n = 20$, $p < 0.001$). Dunn-Bonferroni *post-hoc* tests were conducted to determine the differences between the devices. The results are presented in **Figure 5**. The g.LADYbird device yielded significantly higher SNR values than the Trilobite ($z = -5.409$, $p < 0.001$, $r = 1.2$), EPOC ($z = -6.339$, $p < 0.001$, $r = 1.4$), and Jellyfish devices ($z = -5.832$, $p < 0.001$, $r = 1.3$). The g.SAHARA showed results that were similar to those of g.LADYbird for these three devices (Trilobite: $z = 4.226$, $p < 0.001$, $r = 0.9$; EPOC: $z = -5.155$, $p < 0.001$, $r = 1.2$; Jellyfish: $z = -4.648$, $p < 0.001$, $r = 1.04$). Furthermore, the BR8+ device showed significantly higher SNR values than the EPOC ($z = 3.803$, $p < 0.01$, $r = 0.9$) and Jellyfish devices ($z = -3.296$, $p < 0.05$, $r = 0.7$). All of the obtained effect sizes for the previously mentioned correlation coefficients for device pairs could be interpreted as large according to the guidelines of Cohen (1992).

4.2. Evaluation in Frequency Domain

To evaluate the signal quality in the frequency domain, we conducted a statistical test for each hypothesis. A separate statistical inference evaluation was performed for each device because of the substantial differences between the devices. These arose from the different numbers of electrodes, different electrode layouts, different reference electrodes, and different electrode types. Although those differences did not allow for a statistical inference analysis among the devices, determining a separate inferential statistic for each device seemed to be appropriate to test the hypotheses. The results for the devices could only be compared descriptively. Furthermore, it should be mentioned that evaluations of the third and fifth hypotheses were not possible for the Jellyfish device because of its electrode configuration.

For the third hypothesis, we considered the parietal alpha band power values of the rest measurements with the eyes open and closed. We used the Shapiro-Wilk test to assess whether the alpha band power values of these two rest measurements were normally distributed for each device. This was not the case for the eyes-open parietal alpha band power values of all the devices ($p < 0.05$). Similarly, the alpha band power with the eyes closed was not normally distributed for most of the devices, with the exception of g.SAHARA and g.LADYbird ($p > 0.05$). Hence, for comparison purposes, we conducted a Wilcoxon paired difference test for each EEG system. The results are

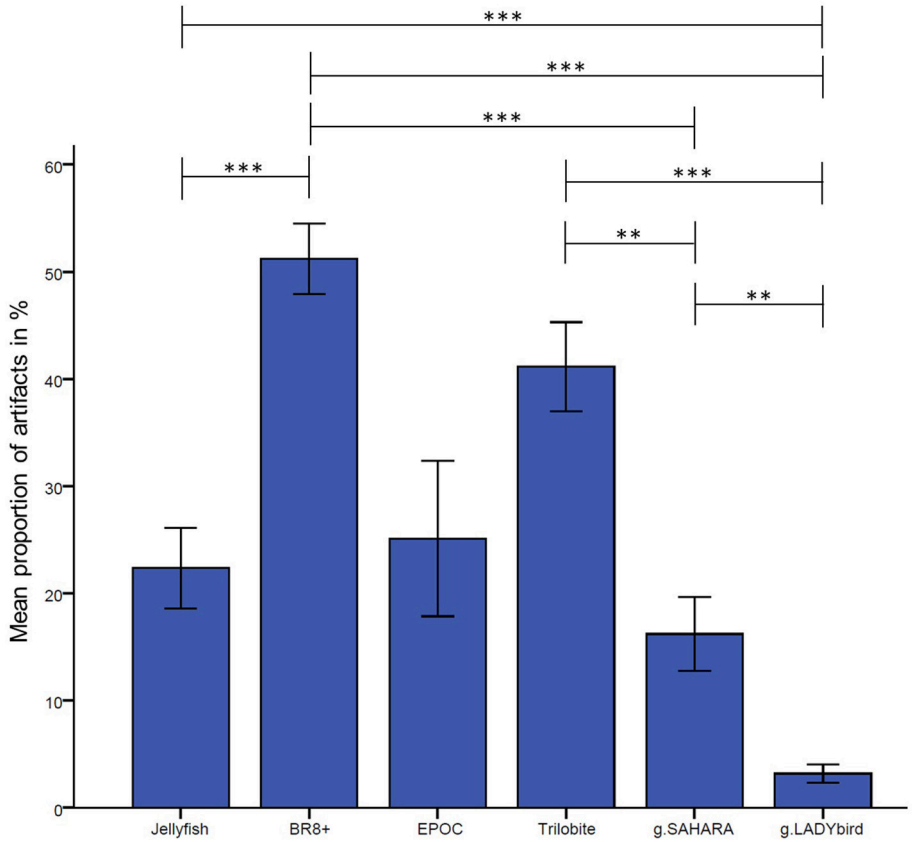


FIGURE 3 | Proportion of manually tagged artifacts in EEG averaged over channels and subjects for each device (calculation of analysis of variance with repeated measures design and Bonferonni-corrected *post-hoc* tests: ***: $p \leq 0.001$; **: $0.001 < p \leq 0.01$; *: $0.01 < p \leq 0.05$; error bars indicate ± one standard deviation).

presented in **Figure 6A**. They show significant differences in the alpha frequency band power values between the eyes open and eyes closed for all of the devices except the Trilobite device ($p = 0.19$).

We used a similar procedure for the fourth hypothesis. Hereby, the theta band power values of the 0-back and stop signal tasks were considered. For all of the devices, the theta band power of the 0-back task was approximately normally distributed, whereas that of the stop signal task was not, as assessed by the Shapiro-Wilk test (Jellyfish and BR8+ with $p < 0.05$). Hence, a Wilcoxon test was conducted. **Figure 6B** shows the results. A significant increase in the frontal theta band power between the easy and more demanding tasks could only be obtained for the Jellyfish and g.LADYbird devices.

Finally, in order to prove our last hypothesis, we examined the alpha band power values of the two cognitive tasks. The Shapiro-Wilk test indicated that during the 0-back task, the alpha band power was not normally distributed for any device ($p < 0.05$). During the stop signal task, the alpha band power was normally distributed for almost all of the devices except the EPOC and g.LADYbird ($p < 0.05$). Thus, a Wilcoxon test had to be applied. The paired difference test between the easy and demanding tasks yielded significant decreases in the parietal alpha band power values for the BR8+, g.SAHARA, and g.LADYbird devices (**Figure 6C**).

5. DISCUSSION AND CONCLUSION

A visual examination of the signals in the time domain and statistical analysis of their proportions of artifacts showed that the gel-based g.LADYbird device had the fewest disturbances, as postulated by hypothesis 1. Among the gel-free devices, the g.SAHARA device had the best performance, with only a small percentage of artifact-contaminated segments. We also want to remind the reader that no significant differences at all could be identified for the EPOC device. This was probably due to the high variance among the subjects and requires a discussion to provide useful information for the use of this device. It is a fact that the headset did not provide a good fit for the various head sizes of the subjects. In these cases, the electrodes did not make good contact with the skin, and the recorded signals included noise interference at 23 and 28 Hz. We assumed that in the case of loose electrode contact, the device caused aliasing artifacts from the electrical mains. Thus, we contacted the manufacturer for a detailed explanation. Their technical support stated that "the

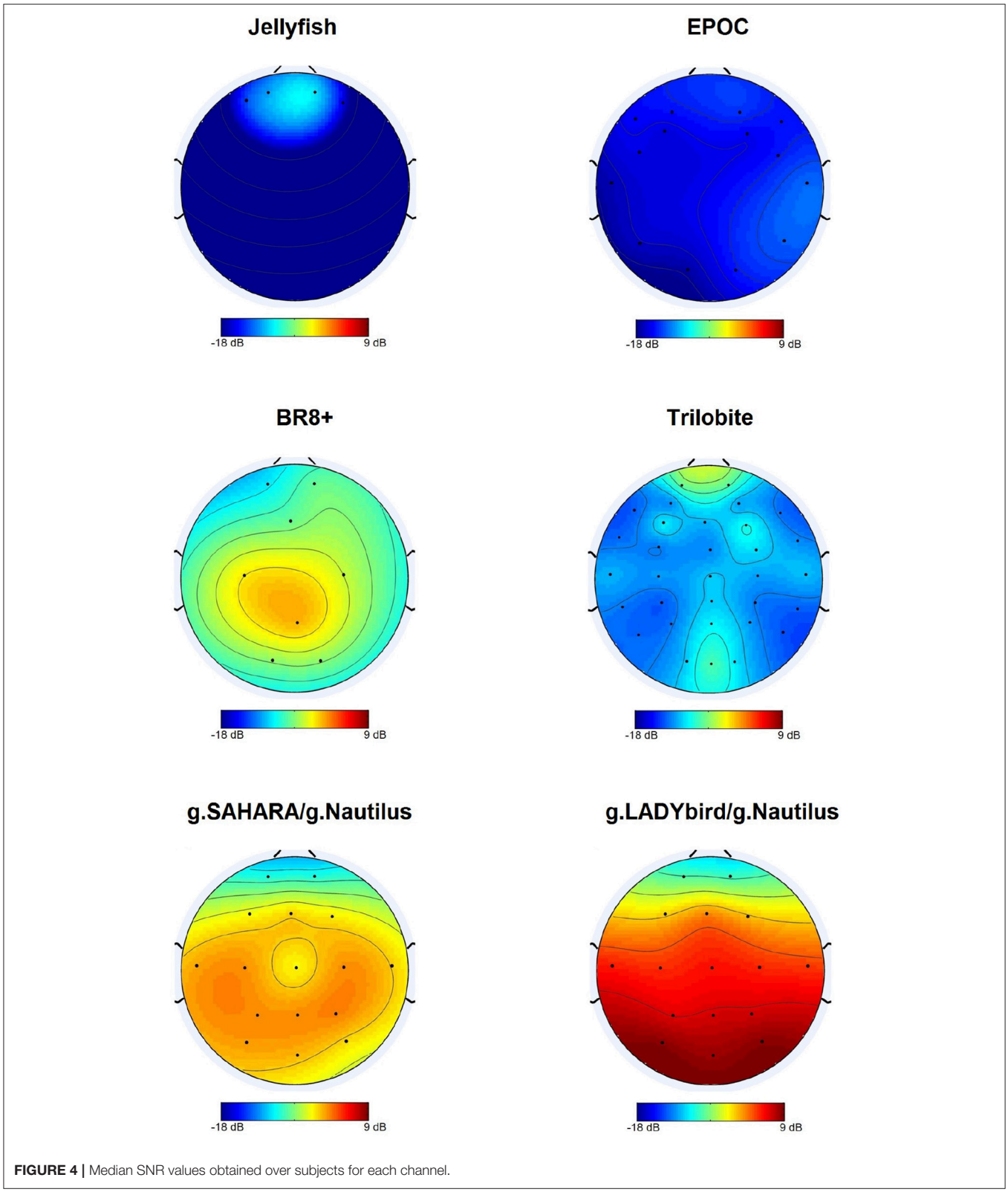


FIGURE 4 | Median SNR values obtained over subjects for each channel.

problem arises because the common mode sense active electrode and driven right leg passive electrode pair cannot cancel the ambient noise, either because the headset is not on a human, or because the connections at the reference locations (behind and 30° above the ears, or directly behind each ear) are not making good contact." We concluded that the variance in the artifact

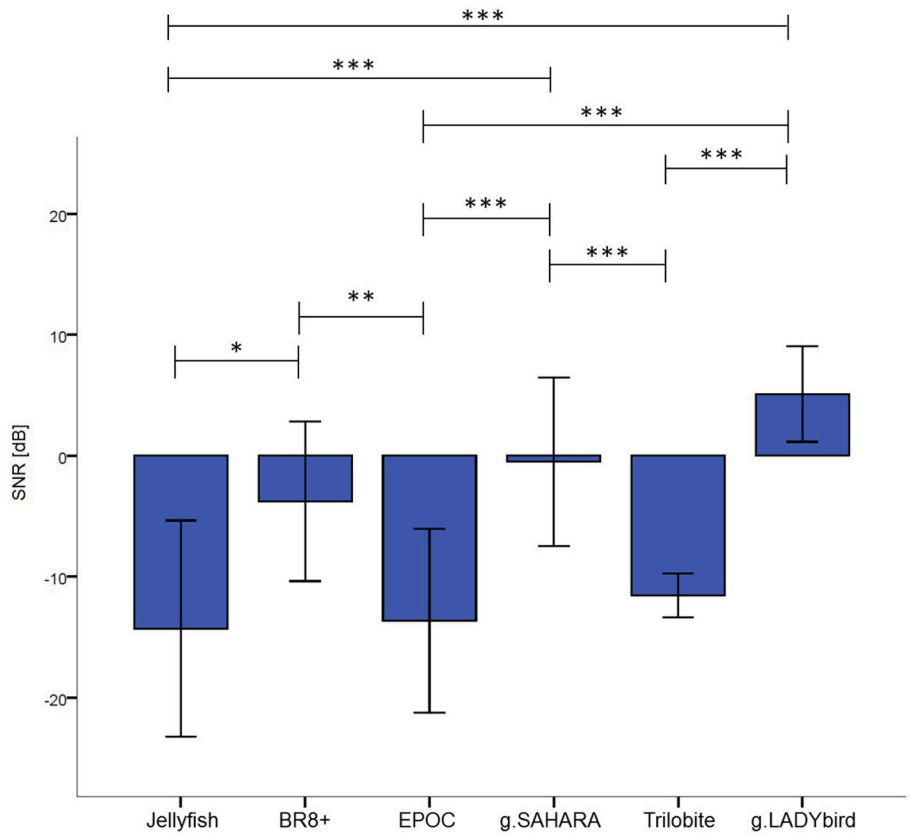


FIGURE 5 | Median SNR values over channels and subjects for each device (calculation of Friedman test of differences and Bonferonni corrected *post-hoc* tests: ***: $p \leq 0.001$; **: $0.001 < p \leq 0.01$; *: $0.01 < p \leq 0.05$; error bars indicate $\pm$ one standard deviation).

proportions among the subjects was large because of the difficulty of adapting the device to the different head sizes. However, the EPOC device is only manufactured in one size, which leads to bad outcomes regarding the signal quality.

For our second hypothesis, we used the signal-to-noise ratio as a criterion to characterize the signal quality of the devices. For all of the devices, the obtained SNR range was quite low, from −18 to 9 dB, and within the range found in the literature. As expected, the SNRs were lower in the frontal areas, which were contaminated by eye artifacts (Goldenholz et al., 2009; Mishra and Singla, 2013; Radüntz et al., 2015). The gel-based g.LADYbird device yielded the best SNR value. A statistical analysis showed that it was significantly higher than the three poorest SNRs of the Trilobite, EPOC, and Jellyfish devices. Among the gel-free devices, we obtained the best SNR value for g.SAHARA. Similar to the values of the g.LADYbird device, g.SAHARA's SNR was significantly higher than the SNR values of the Trilobite, EPOC, and Jellyfish devices. However, remarkably, and in contrast to the g.LADYbird device, none of the gel-free devices could yield SNR values greater than 0 dB (**Figure 5**). This indicated that the ratio between the signal and noise was smaller than one. The noise was superimposed on the signal, which could prove to be particularly problematic in clinical practice, where precise measurements are required.

Our first two hypotheses concentrated on evaluating the EEGs in the time domain. While this evaluation aimed at the first instance to identify the very obvious differences regarding the devices' artifact susceptibility, our evaluation in the frequency domain went a step further. After removing all of the artifact-contaminated segments, we wanted to look deeper at the signal and determine whether it reflected the actual brain activity. For this, we postulated three additional hypotheses based on the well-studied behavior of the EEG. If the devices effectively recorded a brain signal, the Berger effect had to be clearly noticeable. Furthermore, as task demands became greater, we expected an increase in the frontal theta frequency band power and a decrease in the parietal alpha frequency band power.

Significantly, for the gel-based g.LADYbird device, all three frequency-domain hypotheses were proven to be true. For the g.SAHARA and BR8+ devices, significant differences could be obtained regarding the Berger effect and decrease in the parietal alpha band power during the demanding cognitive task. The EPOC device yielded significant differences only for the Berger effect. The Jellyfish device was included only in the examination of the frontal theta band power behavior. It was the only device among the gel-free devices that was able to register a significant increase in the theta band power as task demands increased. Only one device did not show any significant changes in the signal's

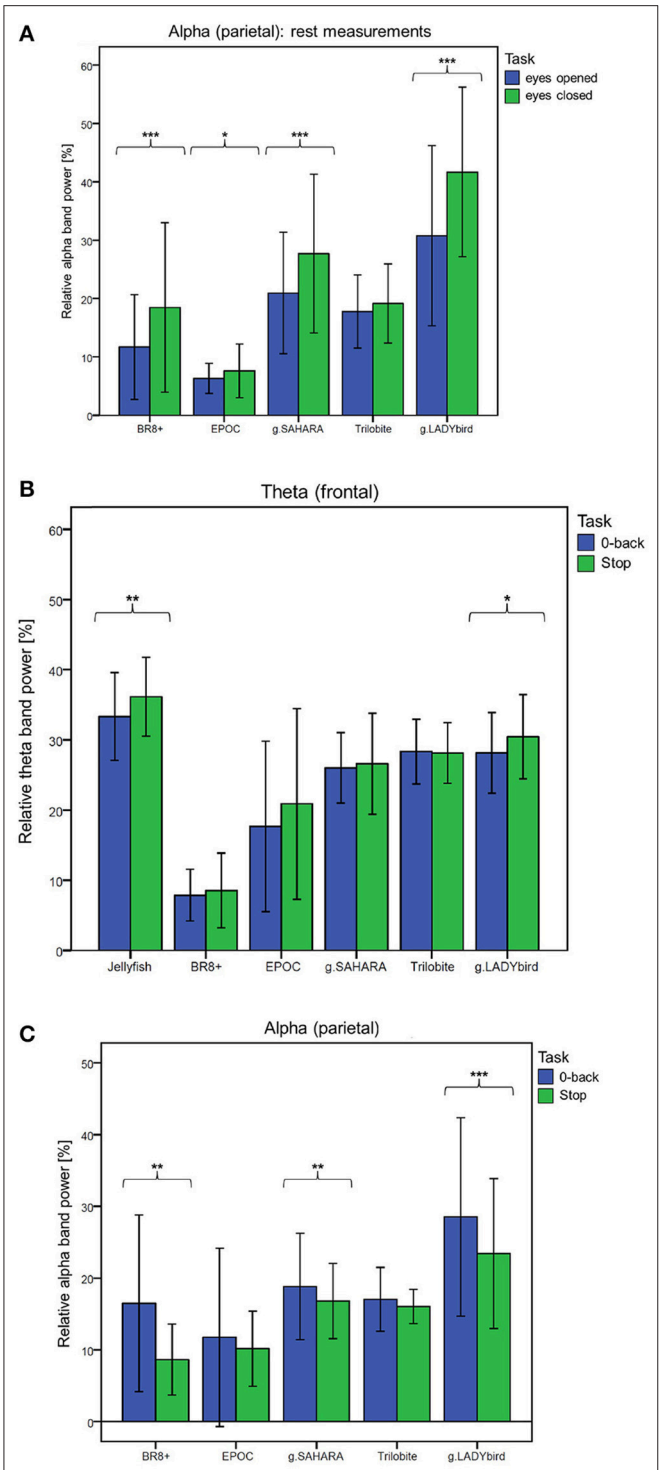


FIGURE 6 | Frequency band differences in respect to different conditions averaged over channels and subjects and considered for each device separately (calculation of Wilcoxon paired difference test for not-normally distributed data; ***: $p \leq 0.001$; **: $0.001 < p \leq 0.01$; *: $0.01 < p \leq 0.05$; error bars indicate ± one standard deviation). **(A)** Hypothesis 1: Behavior of parietal alpha band power during rest measurements with eyes open and eyes closed. **(B)** Hypothesis 2: Behavior of frontal theta band power during easy and more demanding cognitive tasks. **(C)** Hypothesis 3: Behavior of parietal alpha band power during easy and more demanding cognitive tasks.

TABLE 2 | Signal quality results of tested EEG devices (***: $p \leq 0.001$; **: $0.001 < p \leq 0.01$; *: $0.01 < p \leq 0.05$).

Device	EPOC	Trilobite	Jellyfish	BR8+	g.SAHARA	g.LADYbird
Proportion of artifacts [%]	25.11	41.14	22.36	51.22	16.21	3.19
SNR [dB]	−13.66	−11.55	−14.31	−3.78	−0.50	5.09
Berger effect	*		–	***	***	***
Increase in frontal theta			**			*
Decrease in parietal alpha				**	**	***

band power in reference to any of our last three hypotheses: the Trilobite device.

To conclude, all of the devices tested are mobile and do not limit a subject's mobility. All of the devices, except the g.LADYbird device, are easily applicable by the subjects themselves because of their gel-free electrodes. The signal quality results yielded by this study are summarized in **Table 2**. In order to provide useful information to practical users of EEG devices, in the following, we indicate which system could be used under which condition.

Outstanding performances were obtained for the traditional gel-based but mobile g.LADYbird/g.Nautilus device. None of the other emerging devices could reach its signal quality. This device can be recommended for neuroscience research where precise measurements are required.

The signal quality of the g.SAHARA/g.Nautilus device was the best among the gel-free devices and could be considered quite satisfactory. The g.SAHARA/g.Nautilus seems to be a good solution for conducting field experiments. A potential issue could be user acceptance because of the not very flattering cap design and its comfort. A long wearing time for the pin electrodes could be a major problem. Within the framework of our study, we used several questionnaires regarding user experience. The obtained results will be presented in a following paper.

The remaining devices did not meet our requirement of an appropriate signal quality, although some readers could decide to use them for mobile applications.

The EPOC and BR8+ devices suffered from a large proportion of artifacts caused by a poor fit, depending on the subject's head size and form. Hence, they can only be recommended for use if they are guaranteed to perfectly fit the subject's head, e.g., personalized brain-computer applications.

Potential users of the Jellyfish device should be aware that the device only measures the frontal brain activity. In addition, the signal of the frontal electrodes is contaminated by a large number of artifacts. Furthermore, the small number of electrodes does not facilitate the application of artifact-correction algorithms that employ ambient information. However, potential applications suitable for this device could be located in the gaming or bio-feedback sector.

Finally, the results of the Trilobite device were unsatisfactory. This was because of the negative evaluations in both the time domain and frequency domain. A recommendation for the use of the Trilobite device cannot be given based on the obtained results.

It has to be mentioned that the EEG equipment market shows rapid development. During this study, new devices appeared on the market that could not be tested, e.g., the actiCAP Xpress Twist/LiveAmp device by BrainProducts. Furthermore, there is a new highly innovative approach using in-ear EEG technology (Looney et al., 2012; Goverdovsky et al., 2017).

For triggered data from event-related potentials, Oliveira et al. (2016) have already proposed metrics for evaluating new EEG technologies. However, our study design and the proposed method for evaluating the signal quality of devices could easily be used in subsequent studies of new devices and continuous data without triggers. Such a benchmark would allow for the evaluation of further emerging EEG technology and the integration of the test results from new devices into the findings already in existence. This would make it possible to compare emerging EEG devices.

AUTHOR CONTRIBUTIONS

TR initiated the project and was responsible for the overall conception of the investigation and the data analysis. Data interpretation was performed by TR. The manuscript was written by TR.

ACKNOWLEDGMENTS

I would like to thank Friederice Schröder for conducting the experiments, Marion Freyer for the visual inspection of the data and manual artifact marking, and my student assistants Lea Rabe and Emilia Cheladze for their daily operational and computational support. I would also like to thank Gabriele Freude for her general project support. Furthermore, I would like to express my sincere appreciation to Beate Meffert for her valuable and constructive suggestions and her critical editing of the manuscript. More information about the project that acquired our EEG data can be found at http://www.baua.de/DE/Aufgaben/Forschung/Forschungsprojekte/f2402.html.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2018.00098/full#supplementary-material

REFERENCES

Berger, H. (1929). Über das Elektrenkephalogramm des Menschen. *Archiv für Psychiatr. Nervenkrankheiten*, 87, 527–570. doi: 10.1007/BF01797193

Bulea, T. C., Kim, J., Damiano, D. L., Stanley, C. J., and Park, H. S. (2015). Prefrontal, posterior parietal and sensorimotor network activity underlying speed control during walking. *Front. Hum. Neurosci.* 9:247. doi: 10.3389/fnhum.2015.00247

Callan, D. E., Durantin, G., and Terzibas, C. (2015). Classification of single-trial auditory events using dry-wireless eeg during real and motion simulated flight. *Front. Syst. Neurosci.* 9:11. doi: 10.3389/fnsys.2015.00011

Cohen, J. (1992). A power primer. *Psychol. Bull.* 112, 155–159. doi: 10.1037/0033-2909.112.1.155

Debener, S., Minow, F., Emkes, R., Gandras, K., and de Vos, M. (2012). How about taking a low-cost, small, and wireless EEG for a walk? *Psychophysiology* 49, 1449–1453. doi: 10.1111/j.1469-8986.2012.01471.x

Dimoska, A. (2005). *Electrophysiological Indices of Response Inhibition in the Stop-Signal Task*. Ph. D. Thesis, University of Wollongong.

Duvinage, M., Castermans, T., Petieau, M., Hoellinger, T., Cheron, G., and Dutoit, T. (2013). Performance of the emotiv epoc headset for p300-based applications. *Biomed. Eng. OnLine* 12:56. doi: 10.1186/1475-925X-12-56

Forney, E., Anderson, C., Davies, P., Gavin, W., Taylor, B., and Roll, M. (2013). *A Comparison of Eeg Systems for Use in p300 Spellers by Users with Motor Impairments in Real-World Environments*. Graz: Graz University of Technology Publishing House.

Gazzaniga, M., Ivry, R., and Mangun, G. (2013). *Cognitive Neuroscience: The Biology of the Mind*, 4th Edn. New York, NY: W. W. Norton & Company.

Gevins, A., Smith, M. E., Leong, H., McEvoy, L., Whitfield, S., Du, R., et al. (1998). Monitoring working memory load during computer-based tasks with EEG pattern recognition methods. *Hum. Factors* 40, 79–91. doi: 10.1518/001872098779480578

Goldenholz, D. M., Ahlfors, S. P., Hämäläinen, M. S., Sharon, D., Ishitobi, M., Vaina, L. M., et al. (2009). Mapping the signal-to-noise ratios of cortical sources in magnetoencephalography and electroencephalography. *Hum. Brain Mapp.* 30, 1077–1086. doi: 10.1002/hbm.20571

Goverdovsky, V., von Rosenberg, W., Nakamura, T., Looney, D., Sharp, D. J., Papavassiliou, C., et al. (2017). Hearables: multimodal physiological in-ear sensing. *Sci. Rep.* 7:6948. doi: 10.1038/s41598-017-06925-2

Grozea, C., Voinescu, C., and Fazli, S. (2011). Bristle-sensors – low-cost flexible passive dry eeg electrodes for neurofeedback and bci applications. *J. Neural Eng.* 8:025008. doi: 10.1088/1741-2560/8/2/025008

Grummett, T. S., Leibbrandt, R. E., Lewis, T. W., DeLosAngeles, D., Powers, D. M., Willoughby, J. O., et al. (2015). Measurement of neural signals from inexpensive, wireless and dry eeg systems. *Physiol. Meas.* 36:1469. doi: 10.1088/0967-3334/36/7/1469

Guger, C., Krausz, G., Allison, B. Z., and Edlinger, G. (2012). Comparison of dry and gel based electrodes for p300 brain-computer interfaces. *Front. Neurosci.* 6:60. doi: 10.3389/fnins.2012.00060

Johnstone, S. J., Blackman, R., and Bruggemann, J. M. (2012). Eeg from a single-channel dry-sensor recording device. *Clin. EEG Neurosci.* 43, 112–120. doi: 10.1177/1550059411435857

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *J. Exp. Psychol.* 55:352. doi: 10.1037/h0043688

Logan, G. D. (1994). "Chapter On the ability to inhibit thought and action: A users' guide to the stop signal paradigm," in *Inhibitory Processes in Attention, Memory, and Language*, eds D. Dagenbach and T. H. Carr (San Diego, CA: Academic Press), 189–239.

Looney, D., Kidmose, P., Park, C., Ungstrup, M., Rank, M., Rosenkranz, K., et al. (2012). The in-the-ear recording concept: user-centered and wearable brain monitoring. *IEEE Pulse* 3, 32–42. doi: 10.1109/MPUL.2012.2216717

Luu, T. P., Nakagome, S., He, Y., and Contreras-Vidal, J. (2017). Real-time eeg-based brain-computer interface to a virtual avatar enhances cortical involvement in human treadmill walking. *Sci. Rep.* 7, 8895. doi: 10.1038/s41598-017-09187-0

Mishra, P., and Singla, S. K. (2013). Artifact removal from biosignal using fixed point ICA algorithm for pre-processing in biometric recognition. *Meas. Sci. Rev.* 13, 7–11. doi: 10.2478/msr-2013-0001

Mullen, T., Kothe, C., Chi, Y. M., Ojeda, A., Kerth, T., Makeig, S., et al. (2013). "Real-time modeling and 3d visualization of source dynamics and connectivity using wearable eeg," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*(Osaka), 2184–2187.

Mullen, T. R., Kothe, C. A., Chi, Y. M., Ojeda, A., Kerth, T., Makeig, S., et al. (2015). Real-time neuroimaging and cognitive monitoring using wearable dry EEG. *IEEE Trans. Biomed. Eng.* 62, 2553–2567. doi: 10.1109/TBME.2015.2481482

Nikulin, V. V., Kegeles, J., and Curio, G. (2010). Miniaturized electroencephalographic scalp electrode for optimal wearing comfort. *Clin. Neurophysiol.* 121, 1007–1014. doi: 10.1016/j.clinph.2010.02.008

Oliveira, A. S., Schlink, B. R., Hairston, W. D., König, P., and Ferris, D. P. (2016). Proposing metrics for benchmarking novel EEG technologies towards real-world measurements. *Front. Hum. Neurosci.* 10:188. doi: 10.3389/fnhum.2016.00188

Radüntz, T. (2016). *Kontinuierliche Bewertung psychischer Beanspruchung an informationsintensiven Arbeitsplätzen auf Basis des Elektroenzephalogramms*. Ph. D. thesis, Humboldt-Universität zu Berlin, Department of Computer Science(Berlin).

Radüntz, T., Scouten, J., Hochmuth, O., and Meffert, B. (2015). EEG artifact elimination by extraction of ICA-component features using image processing algorithms. *J. Neurosci. Methods* 243, 84–93. doi: 10.1016/j.jneumeth.2015.01.030

Ries, A., Touryan, J., Vettel, J., McDowell, K., and Hairston, W. (2014). A comparison of electroencephalography signals acquired from conventional and mobile systems. *J. Neurosci. Neuroeng.* 3, 10–20. doi: 10.1166/jnsne.2014.1092

Rogers, J. M., Johnstone, S. J., Aminov, A., Donnelly, J., and Wilson, P. H. (2016). Test-retest reliability of a single-channel, wireless eeg system. *Int. J. Psychophysiol.* 106(Suppl. C), 87–96. doi: 10.1016/j.ijpsycho.2016.06.006

Saab, J., Battes, B., and Grosse-Wentrup, M. (2011). "Simultaneous eeg recordings with dry and wet electrodes in motor-imagery," in *12th Conference of Junior Neuroscientists of Tübingen (NeNA 2011)* (Heiligkreuztal).

Sullivan, T. J., Deiss, S. R., and Cauwenberghs, G. (2007). "A low-noise, non-contact eeg/ecg sensor," in *2007 IEEE Biomedical Circuits and Systems Conference* (Montreal, QC), 154–157.

Zander, T. O., Lehne, M., Ihme, K., Jatzev, S., Correia, J., Kothe, C., et al. (2011). A dry eeg-system for scientific research and brain-computer interfaces. *Front. Neurosci.* 5:53. doi: 10.3389/fnins.2011.00053

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Cross-Modality Matching for Evaluating User Experience of Emerging Mobile EEG Technology

Thea Radüntz and Beate Meffert, *Senior Member, IEEE*

***Abstract*—Emerging technology for brain-state monitoring offers the possibility to conduct measurements outside the laboratory. However, user-experience research is lacking. In this article, we present and test an approach for determining the development of user experience in the course of time using the so-called cross-modality matching (CMM). We conducted experiments with 24 subjects and evaluated seven mobile electroencephalography (EEG) devices. Using the CMM method, we registered the headset pressure of the EEG devices and subject's mood. We are able to identify a correlation between headset pressure and mood and to observe time trends. Subjects rated the heaviest, pin-based device as less comfortable in the course of time. The gel-based EEG cap is the most comfortable device regarding its long-time properties. The CMM approach for user-experience evaluation of new EEG technologies is direct, rapid, and easy to perform. This fact creates new opportunities for future studies in the field of user experience and human factors.**

***Index Terms*—Dry sensors, electroencephalography (EEG), psychophysical methods, usability testing and evaluation, wearable devices.**

I. Introduction

REGISTRATION of brain activity by means of electroencephalography (EEG) outside the lab is of increasing interest but also coupled with various challenges. The lack of research about user acceptance regarding the measuring technique is one of them. Meanwhile, mobile and easier to use EEG devices are emerging. They make use of wireless signal transmission and allow the subject to move more freely. Additionally, gel-free sensors enable a quick and easy application of the electrodes. The wearing comfort of the new devices is still unknown, as well as whether user acceptance is improved relative to traditional EEG acquisition. For the use of the devices in future studies, it is of major importance that they do not cause head pressure, discomfort issues, or alter subject's mood state. This is particularly important if the subjects are asked to wear the device for a longer period of time. Knowledge about the wearing time, which is free of complaints or inconvenience can be especially important for many investigations.

Usability studies with more than ten subjects and a within-subject design for the comparison of more than three mobile, consumer-grade EEG devices are rare. Little is known about the evolution of comfort and the influence of the device on subject's mood in the course of time [1]. The few studies involving user-experience research concentrated on one device and used the traditional method of questionnaires to register subjective ratings ([2]–[5]). In the study of Ekandem *et al.* [6] participants were asked to evaluate two devices by completing a post-experiment comfort survey after 15 min of wearing the device. Three different EEG headsets were tested by Nijboer *et al.* [7]. The 13 subjects participating wore every device for approximately an hour during three sessions. At the end of each session, they answered questions regarding the usability of the headset by means of questionnaires. The study by Izdebski *et al.* [1] consisted of two experiments. During the first experiment, four devices were tested by four subjects while during the second experiment three devices were tested by nine subjects. Duration of the sessions varied between one and three hours and the usability was assessed at the end of each session by a questionnaire. Hairston *et al.* [8] conducted a usability research experiment with a wearing time of 60 min and three wireless devices. At the end of the session, participants provided comfort ratings by means of a Likert scale and overall preference ratings based on an ordinal scale. At the end of their article, the authors stated that future studies should include the evolution of the ratings over time and not only the subjective ratings conducted after the experiment. This was a major goal of our study.

For this, we employed the method of cross-modality matching (CMM). The CMM method can be traced back to psychophysical research that aims to describe the relationship between changes in the amplitude of a physical stimulus and the subjective perception of these variations.

To recap, an important psychophysical question is the quantitative relation between a stimulus S and its subjective perception. First relations were found experimentally by Weber in 1864. They were characterized by the so called just-noticeable difference JND that described the smallest change ΔS that could be perceived between two stimuli. In this context, Weber noticed that the greater the initial stimulus S, the larger the difference ΔS needed to distinguish between a first and a second stimulus and that the relation was a constant k (Weber's law)

$$k \sim \frac{\Delta S}{S}. \tag{1}$$

Manuscript received October 29, 2019; revised February 12, 2020; accepted April 5, 2020. This work was supported by the Federal Institute for Occupational Safety and Health under Project F 2402. This article was recommended by Associate Editor D. Wu. *(Corresponding author: Thea Radüntz.)*

Thea Radüntz is with the Unit Mental Health and Cognitive Capacity, Federal Institute for Occupational Safety and Health, 10317 Berlin, Germany (e-mail: raduentz.thea@baua.bund.de).

Beate Meffert is with the Department of Computer Science, Humboldt-Universität zu Berlin, 12489 Berlin, Germany (e-mail: meffert@informatik.hu-berlin.de).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2020.2989380

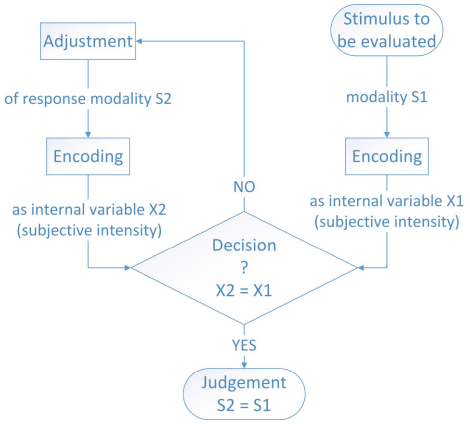


Fig. 1. Principle of CMM according to Sydow and Petzold [31].

Fechner, a scholar of Weber, found that the relation between stimulus S and perception P was logarithmic. In 1957, Stevens introduced an extension ([9]) that showed that sensation magnitude was a power function of stimulus intensity (Stevens' power law)

$$P = b \cdot S^m. \tag{2}$$

Both parameters (the constant b and Stevens' exponent m) are specific for each modality and were already determined for many different ones (e.g., brightness, loudness, apparent length).

Based on the fact, that perceptions of different modalities could be compared to each other, it could be concluded that a stimulus intensity S_1 of one modality could be described by a stimulus intensity S_2 of another modality

$$b_1 \cdot S_1^{m_1} = b_2 \cdot S_2^{m_2} \tag{3}$$

$$\log_{10} S_1 = \frac{m_2}{m_1} \cdot \log_{10} S_2 + \frac{\log_{10} b_2 - \log_{10} b_1}{m_1}. \tag{4}$$

The principle of CMM relies on the idea of perception equalization between different modalities (see Fig. 1). This way, a not measurable modality (e.g., discomfort) can be expressed by a measurable physical modality.

Currently, the method of CMM is gaining again more attention in the scientific community. Researchers show increasing interest to explore [10]–[12] and use the CMM method in order to study human factors and usability aspects [13], [14]. The CMM method can be applied in several situations and research studies. It also provides a good option for conducting ratings from children. The basic idea is similar to a standard procedure used by pediatricians to assess children's pain. They ask the child to press their hand as strong as the pain is. This way, a not measurable modality (e.g., pain, discomfort) can be expressed by a measurable physical modality (e.g., grip force). Application of CMM in user-experience research appears quite appropriate due to the fact that the method can give estimates of sensation's magnitude and, hence, of subjective perception. CMM can be conducted in the course of time and provides real-time measurements. Pepermans and Corlett [15] stated that CMM was well applicable in ergonomics for the evaluation of perceived environmental conditions, which could be difficult to measure subjectively and for the investigation of pain, discomfort, or well-being of a person [16]. At the same time, Pepermans and Corlett [15] conceded that the CMM method has not experienced an extended use in ergonomics. There exist a number of articles employing CMM in order to study somatosensory perception ([17]–[23]), pleasure and pain ([24]–[28]), or discomfort ([29], [30]). The research of Forta *et al.* [30], as one of the latest published articles focusing on practical ergonomics, is the most relevant item to our study. Similar to our aim to assess user experience of several EEG devices, they used CMM for obtaining subject's subjective comfort regarding whole-body vibrations while sitting.

To the best of our knowledge, the CMM method has rarely been used in the context of user-experience research and has never been used for evaluating emerging mobile EEG technology. According to the International Organization for Standardization, user experience is defined as user's perceptions during the use of a product. Thereby, "users' perceptions and responses include the users' emotions, beliefs, preferences, perceptions, comfort, behaviors, and accomplishments that occur before, during, and after use" ([32], Section 3.15). In this study, we focused on two factors of user experience: comfort and mood. We employed the hand-grip force as a modality for CMM ratings. By this, we registered the experienced head pressure caused by the EEG headsets, subject's general mood state, and their change in the course of time. The employment of hand-grip CMM for assessing mood and head pressure evolvement as connected to emerging EEG technology is totally new.

In general, we expected that the wearing time of the devices would have an impact on the comfort. Most of the few studies related to wearing comfort of dry-EEG devices did not explicitly report on the influence of wearing time of the devices (e.g., [3], [33]). The ones that did, reported a duration in the range between 15 and 60 min ([2], [6], [8]). We assumed that subjects' perception of headset pressure would increase after half an hour while their current mood would become worse as long as the headset was worn.

We also addressed the relation between discomfort and mood that had its roots in the research area of embodied cognition, particularly embodied emotion [34]. In this context, the physical condition of a human has a direct influence on the mental state. Hence, we assumed that subject's mood was positively correlated to the wearing comfort of the devices as assessed by the individual perception of head pressure caused by the headset. For the case of a positive correlation, we further assumed that head pressure mediated the relation between device properties and mood. Taken together, we formulated the following hypotheses.

1) During wearing time of the devices, the headset pressure of all EEG devices will increase and subject's mood will get worse, regardless of model, or electrode type.
2) There is a significant positive correlation between current head pressure from the EEG headset and subject's current mood.
   a) The number of electrodes has an effect on the head pressure and, thus, influences subject's mood.
   b) Device's weight has an effect on the head pressure and thus, influences subject's mood.

TABLE I
EEG DEVICES TESTED

Device (manufacturer)	Electrode type	Number of electrodes	Weight
MindCap (mindTec)	Dry sensors	1	119 g
4S Jellyfish (Mindo)	Dry: foam-based	4	95 g
BR8+ (BRI)	Dry: spring-loaded and foam-based	8	269 g
EPOC (Emotive)	Saline based wet sensors	14	116 g
g.SAHARA (g.tec)	Dry pin sensors	16	233 g
g.LADYbird (g.tec)	Gel based	16	165 g
32 Trilobite (Mindo)	Dry: spring-loaded and foam-based	32	524 g

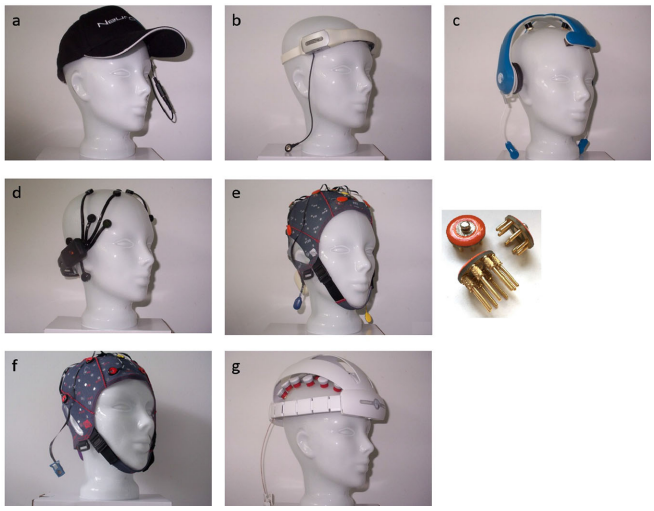


Fig. 2. EEG devices used: (a) MindCap; (b) 4S Jellyfish; (c) BR8+; (d) EPOC; (e) g.SAHARA with dry electrodes; (f) g.LADYbird; and (g) 32 Trilobite.

II. MATERIALS AND METHODS

A. EEG Systems

We conducted market research and chose EEG devices that had left the research-prototype state. These were expected to be suitable for field studies, i.e., quickly and easily applicable without limiting subject's movement while sitting. Seven mobile EEG devices with different characteristics were purchased (see Table I, Fig. 2). In total, six of them were equipped with gel-free electrodes. We also included g.tec's g.LADYbird/g. Nautilus system as a standard gel-based device well suited for mobile use due to its wireless signal transmission and the use of active electrodes.

B. Procedure and Subjects

Our study took place in an office where only the subject and supervisor were present. In total, 24 subjects participating (11 females and 13 males, 26–66 years of age, with a mean age of 42.8) completed in the course of eight consecutive workdays a total of eight sessions with duration of about 90 min each. The first session was aimed at familiarizing the subjects with the method of CMM and the computer tasks they had to perform while wearing the EEG devices. We instructed the subjects that we will not evaluate their performance because the main goal of our study was the evaluation of the devices. During the following sessions, one device per day was selected in

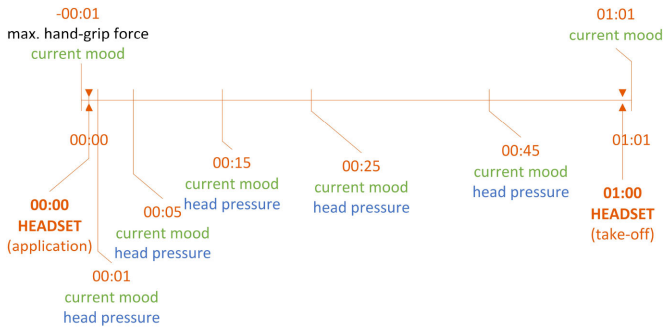


Fig. 3. Timeline of daily sessions for the CMM registration.

(a) (b)

Fig. 4. LabQuest2 interface for data display. (a) Used by the experimenter and hand dynamometer. (b) Used by the subject [36].

random order and tested independently of the others. The study was task independent with strong focus on devices' comfort evaluation. In order to control for side effects related to the tasks, we kept task sequence identical for each device. For more information, we want to draw readers' attention to our paper about the signal-quality evaluation of the devices ([35]). The timeline of the daily session is presented in Fig. 3. The subjects were wearing each device for approximately 60 min. All of the investigations acquired were approved by the local review board of our institution and complied with the tenets of the Declaration of Helsinki. All procedures were carried out with the adequate understanding and written consent of the subjects.

During the following sessions, CMM measurements aimed at evaluating our hypotheses related to the EEG devices. At clearly defined registration time points (see Fig. 3), subjects used the hand-grip force device to answer questions regarding the current head pressure caused by the EEG device and their current mood. They were instructed to apply a greater hand-grip force, the more negative their current mood was. Immediately thereafter, the experienced headset pressure was registered similarly. For doing so, subjects were instructed to grip stronger, the bigger the experienced head pressure of the device was.

Hand-grip force was registered with a strain-gauge-based hand dynamometer by Vernier Company (see Fig. 4). Measurement of hand-grip force was performed with subject's dominant hand and all subjects were right handed. All values were related to subject's maximal grip force that was measured at the beginning of every experimental day.

III. RESULTS

For all subsequent calculations, we used the logarithms of the relative grip-force values and proceeded as described in the

TABLE II
RESULTS OF THE ANOVAS FOR HEADSET PRESSURE AND CURRENT MOOD ASSESSED BY CMM ACROSS REGISTRATION POINTS AND DEVICES

		F	p	η^2
Registration points	Headset pressure	10.083^a	**.001**	.305
	Current mood	8.882^a	**.001**	.279
Device	Headset pressure	6.101	**.001**	.210
	Current mood	1.887^a	.122	.076
Registration points and device	Headset pressure	2.789^a	**.005**	.108
	Current mood	2.150^a	**.039**	.086

Note: Values of .001 are actually $p \leq .001$.
[a] Indicates Mauchly's test of sphericity was significant ($p < .05$) and a Greenhouse–Geisser correction was made to degrees of freedom.

following. All statistical calculations were carried out by means of the SPSS software.

A. Evolvement of Headset Pressure and Mood in the Course of Time

Our first hypothesis assumed that headset pressure would increase in the course of time for all devices and subject's current mood would become worse. We carried out two analyses of variance (ANOVAs) in order to find out if there were significant differences between the registration points, devices, and if there was an interaction between both. The dependent variable was either the head pressure or current mood assessed by CMM. For each ANOVA, we utilized a repeated-measures design with two within-subject factors (seven levels for the device factor and five or seven levels for the registration-points factor for head pressure and current mood, respectively). Results are summarized in Table II. General differences between the levels were examined and tested with posthoc tests (Bonferroni corrected).

CMM values for headset pressure revealed a significant main effect for time, device, and an interaction between both. Fig. 5 (top) shows the headset pressure averaged over the registration points and subjects for each device. Bonferroni corrected posthoc tests showed significant differences between the Trilobite and Jellyfish ($p = .014$), EPOC ($p = .043$), and g.LADYbird ($p = .003$) as well as between the g.LADYbird and BR8+ devices ($p = .014$). The head pressure was increased for the Trilobite device and lowest for the g.LADYbird. Regarding subjects' mood CMM values indicated a significant main effect for time and a weekly significant effect for the interaction between time and device. No significant main effect could be found for the device factor (Fig. 5, bottom). Results for the posthoc tests regarding the registration points and the nature of the interaction between the two factors are shown in Fig. 6 (bottom right) for headset pressure and in Fig. 7 (bottom right) for current mood.

In general, headset pressure decreased five minutes after application of the devices, gradually increased thereafter, and reached its maximum value in 45 min (see Fig. 6). This held true for almost all devices except for the g.SAHARA and g.LADYbird that revealed a flat temporal evolvement. For testing the differences between the registration points for each device separately, we used seven one-factorial, repeated measures ANOVAs with head pressure as dependent variable. Significant differences were obtained only for the Trilobite device ($F(2.44; 56.22) = 17.97$, $p < .001$, $\eta^2 = .439$). Bonferroni corrected posthoc tests revealed a significant difference between the registration point immediately after application of the device and the first 5 min. Thereby, the pressure decreased. Significant changes could also be obtained between the means of the 5th min and the 25th min and between the means of the 45th min and all other registration points before. In these cases, the headset pressure increased significantly in the course of time.

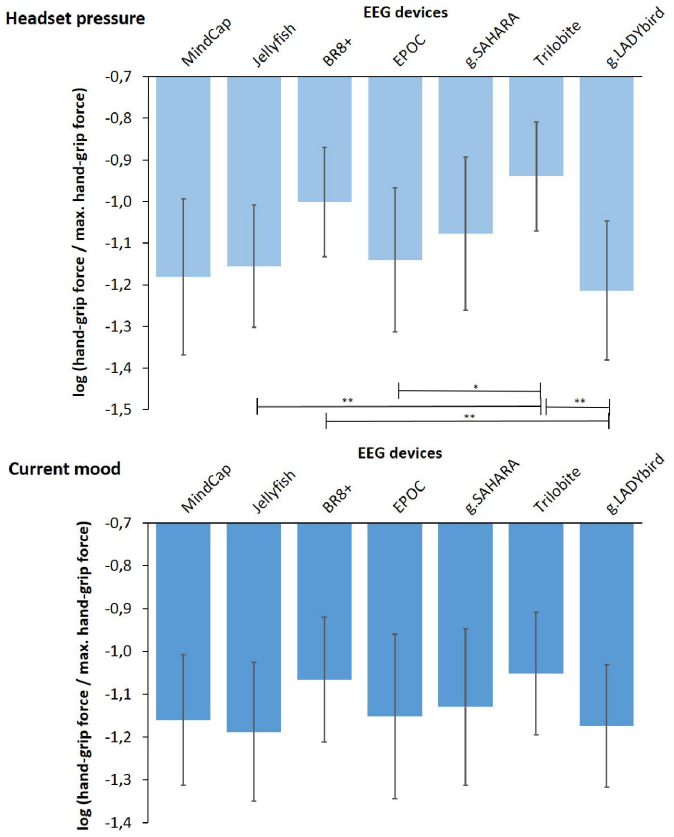


Fig. 5. Headset pressure and current mood as measured by the hand-grip force and averaged over the registration points and subjects for each device (the stronger the grip, the bigger the experienced headset pressure of the device and the more negative the current mood; calculation of analysis of variance with repeated measures design and Bonferonni-corrected posthoc tests: $***$: $p \leq .001$; $**$: $.001 < p \leq .01$; $*$: $.01 < p \leq .05$; error bars indicating the 95% confidence interval).

Descriptive evaluation of subjects' mood revealed that 5 min after device wearing the mood got better for most devices. Thereafter subjects' mood got worse in the course of time and became better after take off of the device. An exception was the g.LADYbird device that did not show the same tendency. In order to statistically evaluate differences between registration points for each device, we computed one-factorial, repeated measures ANOVAs with subjects' current mood as dependent variable. We found significant differences for the devices MindCap ($F(3.92; 90.24) = 2.06$, $p = .09$, $\eta^2 = .082$), BR8+ ($F(2.74; 63.07) = 7.75$, $p < .001$, $\eta^2 = .252$), and Trilobite ($F(2.93; 67.31) = 7.24$, $p < .001$, $\eta^2 = .239$). For the MindCap device Bonferroni corrected posthoc tests showed a significant difference between the registration point after take off of the

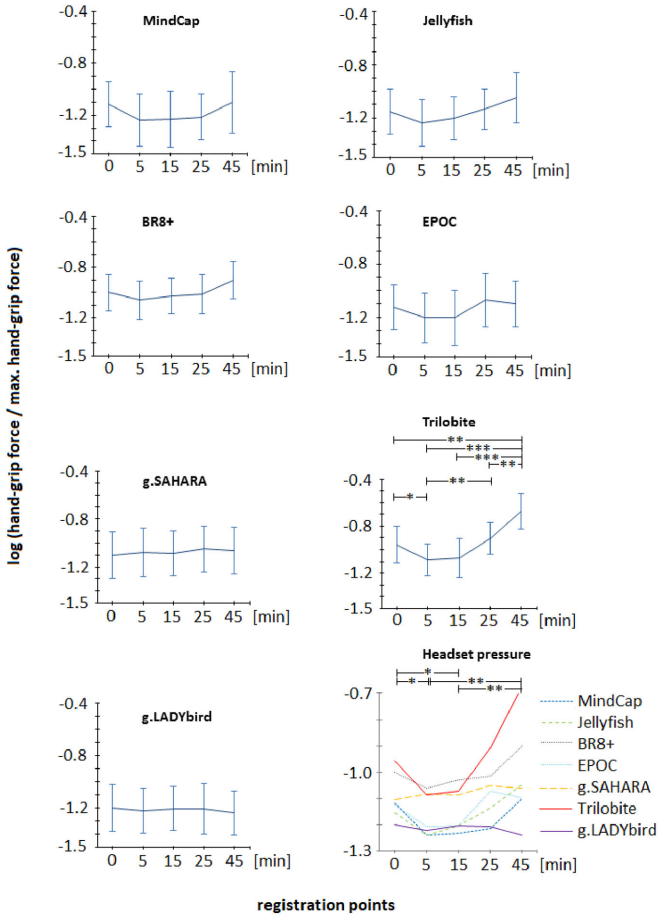


Fig. 6. Headset pressure: Development of the headset pressure measured by the hand-grip force in the course of time and averaged over the subjects for each device and registration point (the stronger the grip, the bigger the experienced headset pressure of the device; calculation of analysis of variance with repeated measures design and Bonferonni-corrected post-hoc tests: $***: p \leq .001$; $**: .001 < p \leq .01$; $*: .01 < p \leq .05$; error bars indicating the 95% confidence interval).

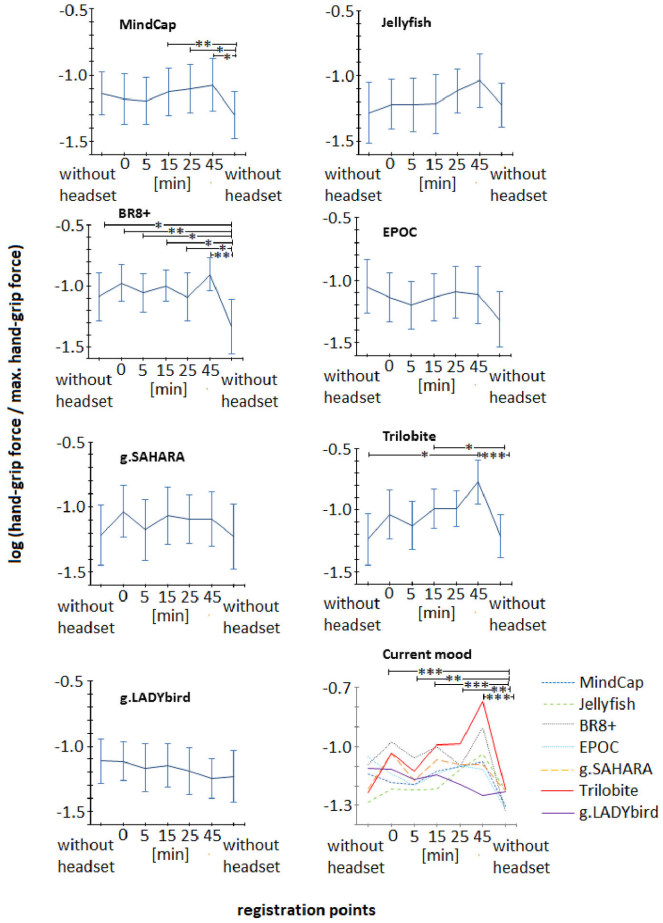


Fig. 7. Current mood: Development of the current mood measured by the hand-grip in the course of time and averaged over the subjects for each device and registration point (the greater the hand-grip force value, the more negative the current mood; calculation of analysis of variance with repeated measures design and Bonferonni-corrected posthoc tests: $***: p \leq .001$; $**: .001 < p \leq .01$; $*: .01 < p \leq .05$; error bars indicating the 95% confidence interval).

headset and the mood in the 15th, 25th, and 45th min. Similar significant differences indicating that the mood got better after take off of the headset were found for the BR8+ device. We observed significant changes in the means between the registration point without the device at the end and all other registration points. Posthoc tests for the Trilobite device revealed that subjects' mood decreased significantly between the registration point before the application of the device and the 45th minute of wearing. Moreover, subjects' mood became significantly better after removal of the device compared to the registration points of the 15th and 45th min, respectively.

B. Correlation Between Headset Pressure and Subject's Mood

We proceeded with the investigation of the relation between subject's mood and experienced head pressure caused by the EEG headsets. To recap, subjects were instructed to apply a greater hand-grip force, the more negative their current mood was. Similarly, they were asked to grip stronger, the bigger the experienced head pressure of the device was. Hence, with increasing hand-grip values for the headset pressure, we expected higher grip-force values for the current mood as well.

TABLE III
CORRELATIONS BETWEEN HEADSET PRESSURE AND SUBJECT'S CURRENT MOOD FOR EACH DEVICE ($N = 24$, **: $p \leq .01$)

	Pearson's correlation coefficient r
MindCap	0.828**
Jellyfish	0.884**
BR8+	0.964**
EPOC	0.772**
g.SAHARA	0.924**
g.LADYbird	0.838**
Trilobite	0.910**

We computed the means for both, the current mood and the headset pressure over subject's single values from the five registration points. This was done for each device separately in order to have an overall value of head pressure and mood from the whole session for each subject and device. In the following, we calculated the correlations between headset pressure and mood for the devices. The results were highly significant, as shown in Table III. All of the obtained effect sizes for the correlation

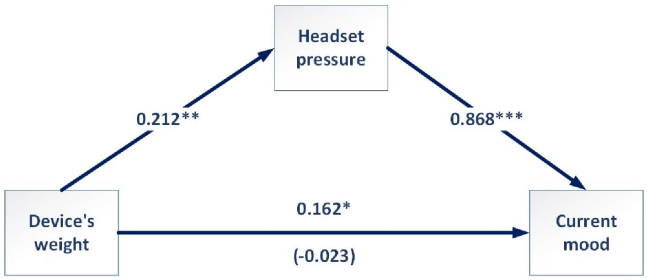


Fig. 8. Standardized regression coefficients for the relationship between device's weight and subject's current mood as mediated by headset pressure. The standardized regression coefficient between device's weight and current mood, controlling for headset pressure, is in parentheses ($***$: $p \leq .001$; $**$: $.001 < p \leq .01$; $*$: $.01 < p \leq .05$; $N = (7 \text{ devices} \times 24 \text{ subjects}) = 168$).

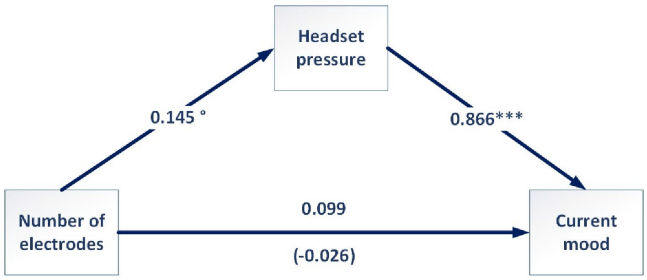


Fig. 9. Standardized regression coefficients for the relationship between number of electrodes and subject's current mood as mediated by headset pressure. The standardized regression coefficient between number of electrodes and current mood, controlling for headset pressure, is in parentheses ($***$: $p \leq .001$; $\circ$: $p = .06$; $N = (7 \text{ devices} \times 24 \text{ subjects}) = 168$).

coefficients of the devices could be interpreted as large according to the guidelines of Cohen denoted for r ([37]).

In the following, we wanted to know if head pressure mediated the relation between device properties and mood. As postulated in our two subhypotheses, the number of electrodes or device's weight could have an effect on the head pressure and, thus, influence subject's mood.

The relationship between device's weight and subject's current mood was mediated by head pressure. As Fig. 8 illustrates, the standardized regression coefficient between device's weight and head pressure was statistically significant, as was the standardized regression coefficient between head pressure and subject's mood. The standardized indirect effect was $0.21 \times 0.87 = 0.18$. Standardized indirect effects were computed for each of 5000 bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th and 97.5th percentiles. The results indicated the indirect coefficient was significant ($b = 0.18$, SE = 0.06, 95% CI = [0.07, 0.29]). Device's weight was no longer a significant predictor of mood after controlling for the mediator head pressure ($b = -0.02$, SE = 0.04, $p = .58$). That is consistent with full mediation.

Results of the investigation of the relationship between number of electrodes and subject's current mood as mediated by head pressure are shown in Fig. 9. The standardized regression coefficient between number of electrodes and head pressure was not significant but near significance level ($p = .06$), while the standardized regression coefficient between head pressure and subject's mood was highly significant. The standardized indirect effect was $0.15 \times 0.87 = 0.13$. Standardized indirect effects were computed for each of 5000 bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th and 97.5th percentiles. The results indicated the indirect coefficient was significant ($b = 0.13$, SE = 0.06, 95% CI = [0.01, 0.25]). There was no significant total effect between number of electrodes and mood ($b = 0.1$, SE=0.08, $p = 0.2$), i.e., number of electrodes did not directly predict subject's mood but only indirectly through head pressure.

IV. Discussion

The main aim of our study was to use the rarely-used method of CMM in order to investigate user-experience issues of emerging EEG technology in the course of time. We employed 24 subjects, tested seven different mobile EEG devices, and conducted ratings of headset pressure and subject's mood by means of CMM.

A. Evolvement of Headset Pressure and Mood in the Course of Time

We expected that in the course of time all EEG headsets would be perceived as burdensome, regardless of model, or electrode type. We also expected that subject's mood would become worse over the headset's wearing time. The hypothesis could not be confirmed for all devices. For headset pressure, we obtained a significant main effect not only regarding registration points but also regarding the devices. Additionally, there was an interaction effect between both. Surprisingly, after 5 min of wearing, headset pressure decreased for almost all devices and became significant for the Trilobite device. Here, we assumed that subjects were familiarized with the new device on their head and the initial discomfort decreased. We have to note that in our study, subjects did not have any previous experience with mobile, dry-electrode EEG devices. Thus, the extent to which the found relations depend on headset experience remains an interesting topic for future research.

In the course of time, the headset pressure generally increased until the 45th minute. This increase was particularly prominent for the Trilobite, our heaviest device. This fits well to our results from the mediator analysis. Furthermore, the Trilobite device showed significant differences regarding head pressure to the devices with soft electrodes (i.e., to the Jellyfish with foam-based electrodes, EPOC with felt-pad electrodes, and g.LADYbird with gel electrodes). For the sake of correctness, we have to mention that these were also the lightest devices. The g.LADYbird device seemed to be the most comfortable device with significant differences to the BR+ and Trilobite devices. It revealed no head-pressure evolvement in the course of time and no significant differences between the registration points. This could be responsible for the highly-significant interaction effect of registration points and device.

Regarding subjects' mood no differences between devices could be obtained although there was a significant interaction between device and registration points. In general, we observed that after the headsets were removed from subject's head the

mood became obviously better for all devices with significant differences to all previous measurements. An exception was observed for the g.LADYbird device where subjects' mood remained almost constant not only in the course of time but also after the removal of the headset. This might be a reason for the weakly-significant interaction effect.

B. Correlation Between Headset Pressure and Subjects' Mood

In our second hypothesis, we suggested a positive significant correlation between headset's comfort and subject's current mood. Correlation analysis yielded large, positive, and significant correlation coefficients for all devices. We concluded that subject's mood was highly correlated to the headset pressure and, thus, to the wearing comfort of the devices.

Furthermore, we investigated if headset pressure mediated the effect of device properties on subject's mood. Results indicated that device's weight was a significant predictor of head pressure and that head pressure was a significant predictor of subject's mood. This supported the mediational hypothesis. After controlling for the mediator, the significant relationship of device's weight and mood became insignificant, indicating a full mediation. This result seemed reasonable because a greater weight could contribute to a greater head pressure and lead to a worse mood.

In contrary, the number of electrodes had only a significant indirect effect through head pressure but no significant total effect on mood. We assumed that there might be other factors apart from the number of electrodes affecting both head pressure and mood (as was the case with device's weight). These might confound the head pressure-mood relationship of our second model. The predictor (in this case the number of electrodes) might be only a part of a more complex model. For instance, the type of electrode could have a greater involvement in subject's current mood than the number of electrodes. This impact could been even amplified by, e.g., device's weight or wearing duration.

Finally, we must be aware that other factors might exist influencing subject's current mood during the sessions. These might be related to the environmental conditions, time on task, or the interaction with the investigator. The randomized testing of the devices across subjects tried to account for some of them. Although, our sample size was relatively large for this kind of study, it was fairly small for elaborate inferential statistics. As a further limitation, we have to mention that during this study new devices appeared on the market, e.g., the actiCAP Xpress Twist/LiveAmp and the saltwater-based electrode system R-Net both by BrainProducts or the new highly innovative approach using in-ear EEG technology ([38], [39]). For evaluating these and further emerging EEG technology, our study design and the proposed CMM method could easily be used. Taken the CMM results as a benchmark to make across-group comparisons [27] would allow for an integration of the test results from new devices into the findings already in existence. This would make it possible to compare emerging EEG devices. Future studies could also evaluate possible effects of specific tasks on user experience of EEG devices as well as further aspects like appealing design, emotions, and pleasure by means of CMM.

V. Conclusion

To sum up, subject's mood and headset pressure were related to each other and changed over the wearing time. This alternation was particularly prominent for the Trilobite device where the changes in the course of time became significant. In contrast, the g.LADYbird device seemed to be the most comfortable. We also found that head pressure was a mediator between device properties and subject's mood, with device's weight as significant predictor. We conclude that developers should attach importance to the weight of the headset for assuring comfort and well-being caused by their devices. In this respect they should be aware of possible interaction effects between the weight, electrode type, and the number of electrodes.

For our investigation, we made use of the method of CMM that is gaining again more attention in the scientific community [40]–[43]. We presented and tested this psycho-physiological approach for evaluating user experience. By this, we compared seven mobile EEG devices and gained reasonable results in the course of time. Although our results might not be surprising, they provide evidence about the feasibility and quality of the CMM ratings. Compared to traditional methods for subjective ratings the CMM approach is direct, rapid, and easy to perform. These facts create new opportunities for future studies in the field of user experience, experimental psychology, and human factors research. Furthermore, our results provide scientific feedback regarding the comfort claims of manufacturers of emerging EEG technology. They are of particular interest for researchers that want to use the new wearable devices for their studies.

In general, CMM offers good possibilities to overcome linguistic or reading barriers or to assess ratings from cognitively impaired subjects. Furthermore, subjects are not limited to a preset scaling or limited number of answers and, thus, less prone to social desirability restrictions caused by predefined answers that could be interpreted as right or wrong. We hope that our article contributes not only to the user-experience evaluation of emerging EEG devices in the course of time but offers also a new example with positive results regarding the applicability of the CMM method.

Acknowledgment

The authors would like to thank G. Menzel and R. Blüthner for their technical and overall support, our student assistant F. Schröder for conducting the experiments, and M. Freyer and the student assistants Y. Cao and I. Pritschke for daily operational support and graphic editing. Furthermore, we would like to thank G. Freude and U. Rose for their general project support. More information about the project where our data were acquired can be found online.[1]

Author contributions: T. Radüntz initiated the project and was responsible for the overall conception of the investigation. Data analysis was performed by T. Radüntz. Data interpretation was performed by T. Radüntz and B. Meffert. The manuscript was written by T. Radüntz. Final critical editing was performed by B. Meffert.

[1][Online]. Available: http://www.baua.de/DE/Aufgaben/Forschung/Forschungsprojekte/f2402.html

Ethics statement: All of the investigations acquired were approved by the local review board of the Federal Institute for Occupational Safety and Health and complied with the tenets of the Declaration of Helsinki. All procedures were carried out with the adequate understanding and written consent of the subjects.

References

[1] K. Izdebski *et al.*, "Usability of EEG systems: User experience study," in *Proc. 9th ACM Int. Conf. Pervasive Technologies Related Assistive Environ.*, 2016, pp. 34-1–34-4. [Online]. Available: http://doi.acm.org/10.1145/2910674.2910714

[2] V. V. Nikulin, J. Kegeles, and G. Curio, "Miniaturized electroencephalographic scalp electrode for optimal wearing comfort," *Clin. Neurophysiology*, vol. 121, pp. 1007–1014, 2010.

[3] C. Grozea, C. Voinescu, and S. Fazli, "Bristle-sensors—low-cost flexible passive dry eeg electrodes for neurofeedback and BCI applications," *J. Neural Eng.*, vol. 8, no. 2, 2011, Art. no. 025008. [Online]. Available: http://stacks.iop.org/1741-2552/8/i=2/a=025008

[4] E. Mathe and E. Spyrou, "Assessment of user experience with a commercial BCI device," in *Proc. Hellenic Conf. Elect. Comput. Eng. Students*, 2015.

[5] T. O. Zander *et al.*, "Evaluation of a dry EEG system for application of passive brain-computer interfaces in autonomous driving," *Frontiers Human Neuroscience*, vol. 11, p. 78, 2017. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnhum.2017.00078

[6] J. Ekandem, T. Davis, I. Alvarez, M. James, and J. Gilbert, "Evaluating the ergonomics of BCI devices for research and experimentation," *Ergonomics*, vol. 55, no. 5, pp. 592–598, 2012. [Online]. Available: http://dx.doi.org/10.1080/00140139.2012.662527

[7] F. Nijboer, B. van de Laar, S. Gerritsen, A. Nijholt, and M. Poel, "Usability of three electroencephalogram headsets for brain-computer interfaces: A within subject comparison," *Interacting Comput.*, vol. 27, no. 5, pp. 500–511, 2015. [Online]. Available: https://doi.org/10.1093/iwc/iwv023

[8] W. Hairston *et al.*, "Usability of four commercially-oriented EEG systems," *J. Neural Eng.*, vol. 11, no. 4, 2014, Art. no. 046018. [Online]. Available: http://stacks.iop.org/1741-2552/11/i=4/a=046018

[9] S. Stevens, "On the psychophysical law," *Psychological Rev.*, vol. 64, no. 3, pp. 153–181, 1957.

[10] T. M. Benz and V. Nitsch, "Is cross-modal matching necessary? A bayesian analysis of individual reference cues," in *Haptics: Science, Technology, and Applications*, D. Prattichizzo, H. Shinoda, H. Z. Tan, E. Ruffaldi, and A. Frisoli, Eds. Berlin, Germany: Springer, 2018, pp. 15–26.

[11] K. M. Gomes and S. L. Riggs, "The effect of age on crossmodal matching using auditory frequency," in *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, vol. 61, no. 1, pp. 1552–1556, 2017. [Online]. Available: https://doi.org/10.1177/1541931213601752

[12] K. Gomes and S. L. Riggs, "Crossmodal matching: A comparison of two methods," in *Proc. Human Factors Ergonom. Soc. Annu. Meeting*, vol. 60, no. 1, pp. 1595–1599, 2016. [Online]. Available: https://doi.org/10.1177/1541931213601368

[13] M. C. Lee and J. Park, "There is no perfect evaluator: An investigation based on prospect theory," *Human Factors Ergonom. Manuf. Service Industries*, vol. 28, no. 6, pp. 383–392, Jun. 2018.

[14] S. L. Riggs and N. Sarter, "Tactile, visual, and crossmodal visual-tactile change blindness: The effect of transient type and task demands," *Human Factors*, vol. 61, no. 1, pp. 5–24, 2019. [Online]. Available: https://doi.org/10.1177/0018720818818028

[15] R. G. Pepermans and E. N. Corlett, "Cross-modality matching as a subjective assessment technique," *Appl. Ergonom.*, vol. 14, no. 3, pp. 169–176, 1983. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0003687083900789

[16] A. Campbell, "Subjective measures of well-being," *Amer. Psychologist*, vol. 31, pp. 117–124, 1976.

[17] L. Walker and P. Walker, "Cross-sensory mapping of feature values in the size-brightness correspondence can be more relative than absolute," *J. Exp. Psychol., Human Perception Perform.*, vol. 42, no. 1, pp. 138–150, 2016.

[18] T. N. Coates, "Source localization in cross modality matching of brightness and loudness of young adults," Communication Disorders, Master's Thesis, David O. McKay School of Education, Provo, UT, USA, 2015.

[19] J. Hayes, "Measuring sensory perception in relation to consumer behavior," in *Rapid Sensory Profiling Techniques*. New York, NY, USA: Elsevier, 2015, pp. 53–69.

[20] S. Simonetti, J. Kim, and C. Davis, "Cross-modality matching of linguistic and emotional prosody," in *Proc. Interspeech*, 2015, pp. 56–59.

[21] J. Venrooij *et al.*,"Perception-based motion cueing: Validation in driving simulation," in *Proc. Eur.: Driving Simul. Conf. Exhib.*, 2015, pp. 153–161.

[22] M. W. Heft and M. E. Robinson, "Age differences in suprathreshold sensory function," *AGE*, vol. 36, no. 1, pp. 1–8, Apr. 2013.

[23] K. Bronner, K. Frieler, H. Bruhn, R. Hirt, and D. Piper, "What is the sound of citrus?," in *Proc. 12th Int. Conf. Music Perception Cognition*, 2012, pp. 142–148.

[24] M. W. Heft and M. E. Robinson, "The challenge of measuring pain in humans," in *Biological Measures of Human Experience Across the Lifespan*. Berlin, Germany: Springer, 2016, pp. 105–115.

[25] C. Morewedge, *Utility: Anticipated, Experienced, Remembered*, 2nd ed. Hoboken, NJ, USA: Wiley, 2016, ch. 10.

[26] G. Lathwal, I. K. Pandit, N. Gugnani, and M. Gupta, "Efficacy of different precooling agents and topical anesthetics on the pain perception during intraoral injection: A comparative clinical study," *Int. J. Clin. Pediatric Dentistry*, vol. 8, pp. 119–122, 2015.

[27] L. Bartoshuk, "The measurement of pleasure and pain," *Perspectives Psychological Sci.*, vol. 9, no. 1, pp. 91–93, Jan 2014.

[28] O. C. G. Colhado *et al.*,"Evaluation of low back pain: Comparative study between psychophysical methods," *Pain Medicine*, vol. 14, no. 9, pp. 1307–1315, Sep. 2013.

[29] Y. Huang and M. J. Griffin, "The discomfort produced by noise and whole-body vertical vibration presented separately and in combination," *Ergonomics*, vol. 57, no. 11, pp. 1724–1738, Aug. 2014.

[30] N. Forta, M. Schust, P. von Löwis, H. Kaiser, and A. Kreisel, "Subjective comfort of fore-and-aft whole-body vibration measured with the cross-modality matching method," in *Proc. 47th U.K. Conf. Human Response Vibration*, Southampton, England, 2012.

[31] H. Sydow and P. Petzold, *Mathematische Psychologie: Mathematische Modellierung und Skalierung in der Psychologie*. New York, NY, USA: Springer, 1982.

[32] EN ISO-9241-210:2019, *Ergonomics of Human-System Interaction—Part 210: Human-Centred Design for Interactive Systems*, Beuth-Verlag, International Organization for Standardization, Geneva, Switzerland, 2019.

[33] A. Pinegger, S. C. Wriessnegger, J. Faller, and G. R. Müller-Putz, "Evaluation of different EEG acquisition systems concerning their suitability for building a brain–computer interface: Case studies," *Frontiers Neuroscience*, vol. 10, p. 441, 2016. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2016.00441

[34] T. F. Price and E. Harmon-Jones, "Embodied emotion: The influence of manipulated facial and bodily states on emotive responses," *Wiley Interdisciplinary Rev., Cogn. Sci.*, vol. 6, no. 6, pp. 461–473, Sep. 2015.

[35] T. Radüntz, "Signal quality evaluation of emerging EEG devices," *Frontiers Physiol.*, vol. 9, p. 98, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fphys.2018.00098

[36] Vernier. (2017) Software and technology. Accessed: Jul. 2, 2017. [Online]. Available: https://www.verniers.com

[37] J. Cohen, "A power primer," *Psychological Bull.*, vol. 112, no. 1, pp. 155–159, 1992.

[38] V. Goverdovsky *et al.*, "Hearables: Multimodal physiological in-ear sensing," *Scientific Rep.*, vol. 7, no. 1, Jul. 2017, doi : 10.1038/s41598-017-06925-2.

[39] D. Looney *et al.*,"The in-the-ear recording concept: User-centered and wearable brain monitoring," *IEEE Pulse*, vol. 3, no. 6, pp. 32–42, Nov. 2012.

[40] S. L. Riggs and N. Sarter, "Crossmodal matching: The case for developing and employing a valid and feasible approach to equate perceived stimulus intensities in multimodal research," *Human Factors*, vol. 61, no. 1, pp. 29–31, 2019. [Online]. Available: https://doi.org/10.1177/0018720818816439

[41] G. Hamilton-Fletcher, K. Pisanski, D. Reby, M. Stefańczyk, J. Ward, and A. Sorokowska, "The role of visual experience in the emergence of cross-modal correspondences," *Cognition*, vol. 175, pp. 114–121, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0010027718300593

[42] B. Pitts, S. L. Riggs, and N. Sarter, "Crossmodal matching: A critical but neglected step in multimodal research," *IEEE Trans. Human-Mach. Syst.*, vol. 46, no. 3, pp. 445–450, Jun. 2016.

[43] R. S. Schaefer, L. J. Beijer, W. Seuskens, T. C. M. Rietveld, and M. Sadakata, "Intuitive visualizations of pitch and loudness in speech," *Psychonomic Bull. Rev.*, vol. 23, no. 2, pp. 548–555, Apr. 2016. [Online]. Available: https://doi.org/10.3758/s13423-015-0934-0

Original Paper

User Experience of 7 Mobile Electroencephalography Devices: Comparative Study

Thea Radüntz[1], Dr rer nat; Beate Meffert[2], Dr-Ing

[1]Mental Health and Cognitive Capacity, Federal Institute for Occupational Safety and Health, Berlin, Germany

[2]Department of Computer Science, Humboldt-Universität zu Berlin, Berlin, Germany

Corresponding Author:
Thea Radüntz, Dr rer nat
Mental Health and Cognitive Capacity
Federal Institute for Occupational Safety and Health
Nöldnerstr 40-42
Berlin, 10317
Germany
Phone: 49 30 51548 4418
Email: raduentz.thea@baua.bund.de

Abstract

Background: Registration of brain activity has become increasingly popular and offers a way to identify the mental state of the user, prevent inappropriate workload, and control other devices by means of brain-computer interfaces. However, electroencephalography (EEG) is often related to user acceptance issues regarding the measuring technique. Meanwhile, emerging mobile EEG technology offers the possibility of gel-free signal acquisition and wireless signal transmission. Nonetheless, user experience research about the new devices is lacking.

Objective: This study aimed to evaluate user experience aspects of emerging mobile EEG devices and, in particular, to investigate wearing comfort and issues related to emotional design.

Methods: We considered 7 mobile EEG devices and compared them for their wearing comfort, type of electrodes, visual appearance, and subjects' preference for daily use. A total of 24 subjects participated in our study and tested every device independently of the others. The devices were selected in a randomized order and worn on consecutive day sessions of 60-min duration. At the end of each session, subjects rated the devices by means of questionnaires.

Results: Results indicated a highly significant change in maximal possible wearing duration among the EEG devices (χ^2_6=40.2, n=24; P<.001). Regarding the visual perception of devices' headset design, results indicated a significant change in the subjects' ratings (χ^2_6=78.7, n=24; P<.001). Results of the subjects' ratings regarding the practicability of the devices indicated highly significant differences among the EEG devices (χ^2_6=83.2, n=24; P<.001). Ranking order and posthoc tests offered more insight and indicated that pin electrodes had the lowest wearing comfort, in particular, when coupled with a rigid, heavy headset. Finally, multiple linear regression for each device separately revealed that users were not willing to accept less comfort for a more attractive headset design.

Conclusions: The study offers a differentiated look at emerging mobile and gel-free EEG technology and the relation between user experience aspects and device preference. Our research could be seen as a precondition for the development of usable applications with wearables and contributes to consumer health informatics and health-enabling technologies. Furthermore, our results provided guidance for the technological development direction of new EEG devices related to the aspects of emotional design.

(JMIR Mhealth Uhealth 2019;7(8):e14474) doi: 10.2196/14474

KEYWORDS

wearable devices; user experience; electroencephalography; mobile applications; electrodes; dry electrodes

Introduction

User Experience Research of Emerging Electroencephalography Technology

In the previous years, registration of brain activity has become more and more popular not only in science but also in the home and gaming sector. Users look forward to identifying and quantifying their mental state directly there where human information processing takes place, and electroencephalography (EEG) offers a way to assess the levels of fatigue, stress, or emotions. The state feedback can then be used to prevent undesired situations, enhance wanted effects, or control devices. The increasing number of publications related to brain-computer interfaces [1-7] indicates an ever-growing interest in communication systems where encoded brain activity from the user is used as an alternative channel to send information to a computer. In addition, progress in sensor technology enables the production of low-cost, light-weighted, and marketable devices. However, extended use of the EEG is hampered by user experience challenges and user acceptance issues regarding the measuring technique.

Only a few years ago, one of the main issues was the limited mobility of the subjects because of the wired connections going from the electrode cap to an amplifier and computer. Meanwhile, wireless signal transmission helps to overcome this problem and allows subjects to move more freely. Further concerns are related to the application of gel electrodes and skin preparation for reducing the impedance. Emerging sensor technology uses gel-free sensors to enable a quick and easy application of the electrodes by the users themselves. For assuring an acceptable signal quality, impedance between electrodes and skin must be low, that is, electrodes need a good and permanent contact to the skin. This becomes particularly difficult to achieve for dry electrodes that work without the conductive gel. Given this, the question of wearing comfort and user experience becomes even more evident.

Finally, there are also user experience issues related to the unflattering visual appearance of the traditional EEG caps and thus linked to the research field of emotional design [8]. The core idea thereby is that products' design strives to elicit positive emotions and thus influence users' perception to provide a greater level of user experience. The 3-level model of emotional design includes the visceral, the behavioral, and the reflective level [8,9]. The visceral is the most basic, immediate level and addresses our first reactions to visual or sensory aspects (eg, aesthetics and quality) of the product. The behavioral level refers to usability aspects of the product, whereas the reflective level comprises conscious cognition. More general, the reflective level asks how well the product fits in with user's current self-image and addresses not only mental and emotional but also social aspects.

To recap, there is growing interest among users in brain state monitoring and increased efforts by developers for developing mobile EEG devices. However, serious user experience research in this field is rare, and it remains still unclear whether user acceptance of the new devices is improved compared with traditional EEG technology. In our study, we aimed to address this issue and advance the state of the art regarding user experience of emerging EEG devices. Thereby, we focused on the wearing comfort of the devices and aspects of emotional design, particularly the behavioral and reflective levels.

Related Work

During the previous years, the advances in sensor technology promoted the research regarding the usability of emerging EEG devices. Most of the published papers concentrated only on device functionality and signal quality comparison between the traditional gel-based electrodes and the new dry electrodes [7,10-12].

Only a small number of studies were concerned with devices' wearing comfort and design requirements. Nikulin et al [13] reported that for designing a new kind of electrodes, they considered not only signal quality but also electrodes' visual appearance and wearing comfort. They put effort to create extremely light and small electrodes that could be applied with some conductive gel directly on the head without any cap or headset. During the study, subjects reported that the electrodes were not noticeable and also not visually detectable by other people. Subjects felt less watched and thus better. Nikulin et al argued that this was particularly important when working outside the laboratory, and subjects were asked to behave naturally and free, in particular, during field experiments in real work environments. However, the main limitation was that the electrodes had to be applied with gel. This application procedure was time consuming and required specific knowledge about electrodes' precise positions on the head. Hence, it had to be done by an experienced investigator and could not be done by the subject itself. A further limitation was that the subjects did not have the opportunity to compare the new electrode device with another.

Similarly, Grozea et al [14] reported on their work on new electrodes with fine, flexible, and metal-coated polymer bristles. The bristles should allow for a good contact through the hair, and simultaneously, they should be comfortable during wearing. The researchers tested the electrodes on subjects (ie, colleagues) that had previous experience with other kinds of electrodes (eg, gel-based and pin electrodes). The subjects concluded that although the bristles electrodes were better than the pin electrodes, the bristles could have been softer and more flexible to increase comfort. Limitations of the study were the small number of subjects participating and the lack of direct comparison among the different kinds of electrodes instead of recalling the wearing comfort from previous experiences.

Comparison studies among different commercial EEG devices regarding user experience were rare. A study by Ekandem et al [15] dealt with the comparison between Emotiv's EPOC device and NeuroSky's MindWave device. Research questions concerned the wearing comfort, the preparation, and the application time. The latter was less than 5 min for both devices and thus clearly less compared with traditional EEG devices. After 15 min of wearing, subjects were asked to answer questions about the overall comfort of the worn device, the length of time they would be able to wear it, and the type of discomfort [15]. Thereby, the EPOC device was rated more comfortable compared with the MindWave device. A main

limitation of the study concerned the wearing time of 15 min because this could be insufficient for determining discomfort issues.

A study by Izdebski et al [16] was divided into 2 similar experiments that tested in total 7 devices. Of 7 devices, 4 devices (g.tec's g.SAHARA, Emotiv's EPOC, ANT Neuro's asalab, and Brain Products' [Brain Products GmbH] actiCAP) were tested by 4 subjects, and the remaining 3 devices (BioSemi's ActiveTwo, Cognionics' Dry System, and Cognionics' Wet System) were tested by 9 subjects. Duration of the sessions varied between 1 and 3 hours, and the usability was assessed at the end of each session by a questionnaire. Surveyed usability aspects were comfort, cap fit, mood, and movement restriction. Izdebski et al reported that the gel-based electrode headsets asalab and actiCAP induced general discomfort although participants did not report an unpleasant feeling under the cap nor a high pressure of the electrodes. Regarding cap fit, the ActiveTwo and systems without adjustment possibilities received negative ratings. The EPOC, g.SAHARA, and asalab devices yielded a more negative mood at the end of the session, whereas the wired systems asalab and actiCAP were rated as more movement restricting. A limitation of the study concerns the lack of a consistent within-subject design and the very different session durations.

Hairston et al [17] conducted a usability research experiment with a wearing time duration of 60 min. They compared 4 EEG devices: 3 wireless EEG systems (Emotiv's EPOC, Advanced Brain Monitoring's B-Alert X10, and QUASAR's HMS) and 1 wired, laboratory-grade device (Bio-Semi's ActiveTwo). The main user experience aspects they focused on, besides signal quality issues, were the adaptability of the devices to different head sizes, comfort, and subjects' device preference. They found that subjects preferred the B-Alert X10 device more than the other 2 wireless systems although it had gel-based electrodes. Subjects reported that the gel-infused pads of the B-Alert X10 device were more comfortable than the others. Finally, Hairston et al stated that future work was needed to systematically study usability factors and improve development efforts of new systems.

To compare the usability of a brain-computer interface for communication, Nijboer et al [18] tested 3 different EEG headsets (g.tec's g.SAHARA, Emotiv's EPOC, and BioSemi's ActiveTwo). Apart from signal quality, Nijboer et al also assessed the speed and ease of headset's setup, subjects' rating about their appearance with headset, comfort, and general device preference. Nijboer et al obtained the highest setup time for the gel-based ActiveTwo device, the best aesthetic ratings for the EPOC device, and the best comfort ratings for the gel-based ActiveTwo and pin-based g.SAHARA devices. Although the EPOC device yielded the worst ratings regarding comfort, it was the device of choice in the ranking of preference. Nijboer et al assumed that aesthetics and ease of use could be more important factors than comfort when it comes to preference ranking. They stated that more research was needed to understand which user experience aspects influence subjects' preference choice.

Table 1 summarizes the above-mentioned studies in a symmetric presentation style. To conclude, considering that duration of registration sessions and thus device wearing can take a long time, comfort requirements are particularly important. Existing studies regarding the usability of EEG headsets indicated that for assuring user acceptance, devices should be lightweight, comfortable, not painful to wear, and with an unobtrusive design. However, limitations of these studies were a limited number of participants, lack of comparisons among different devices, or a too short wearing duration of the EEG headsets. Most of the studies focused primarily on wearing comfort and neglected user experience aspects such as emotional design. In our study, we considered these things and systematically compared 7 different EEG devices.

Table 1. Literature review regarding user experience of emerging electroencephalography technology.

Reference	Devices tested	Electrode type and number	Set size	Wearing duration	User aspects and items	Results
Nikulin et al 2010 [13]	Proprietary development, traditional EEG[a] cap	Miniaturized C-electrodes with gel, 3; standard electrodes with gel, 3	4 subjects	40-60 min	Wearing comfort, tactile sensation, shame	No tactile sensations associated with C-electrode wearing, no negative emotional impact in the presence of others, and no discomfort
Grozea et al [14]	Proprietary development	Dry bristle electrodes; no information about number of electrodes	8 colleagues (2 of them excluded)	<1 hour	Comfort issues	Most subjects reported them to be more advanced than the previously known
Ekandem et al [15]	Emotiv's EPOC, NeuroSky's MindWave	Saline-based, 14; dry, 1	13 subjects (2 of them excluded)	15 min	Comfort and wearing duration	EPOC more comfortable; at least 20 min possible
Izdebski et al [16]	g.tec's g.SAHARA, Emotiv's EPOC, Cognionics' Dry System, ANT Neuro's asalab, Brain Products' actiCAP, BioSemi's ActiveTwo, and Cognionics' Wet System	Dry, 32; saline-based, 14; dry, 64; gel, 128; gel, 64; gel, 128; gel, 64	4 subjects (g.SAHARA, EPOC asalab, and actiCAP); 9 subjects (ActiveTwo, Cognionics' Dry System, and Cognionics' Wet System)	4 subjects (2-3 hours); 9 subjects (1-2 hours)	Comfort, cap fit, mood, and movement restriction	asalab and actiCAP induced general discomfort although participants did not report unpleasant feeling under cap nor high pressure of electrodes; ActiveTwo and systems without adjustment possibilities received negative ratings regarding cap fit; EPOC, g.SAHARA, and asalab yielded a more negative mood at the end of the session; the wired systems asalab and actiCAP were rated as more movement restricting
Hairston et al [17]	Emotiv's EPOC, Advanced Brain Monitoring's B-Alert X10, QUASAR's HMS, and BioSemi's ActiveTwo	Saline-based, 14; gel, 9; dry, 9; gel, 64	16 subjects (3-4 of them excluded)	60 min	Comfort, preference	Most preferred: B-Alert; comfortable to wear
Nijboer et al [18]	g.tec's g.SAHARA, Emotiv's EPOC, BioSemi's ActiveTwo	Dry, 8; saline-based, 14; gel, 32	13 subjects	~1 hour	Speed and ease of setup, appearance with headset, comfort, and general preference	Highest setup time for ActiveTwo; best aesthetic ratings for EPOC; best comfort ratings for ActiveTwo and g.SAHARA; in general, most preferred: EPOC

[a]EEG: electroencephalography.

Research Objectives

As the registration of brain activity outside the laboratory becomes more popular, aspects of user experience attract more attention when new devices are to be developed. Apart from improving wearing comfort that is crucial regarding user experience, developers also put more emphasis on the headset design of the EEG devices. This can lead to extraordinary designs that are not always flattering and easy to use for the user. In such cases, the visual appearance and behavior of the device can influence the well-being of a person [13].

Our first research objective was concerned with the test of the devices. First, we referred to the well-known issue of wearing comfort linked to the different electrode types and the question of how comfortable the different electrodes were after a longer wearing time. We assumed that maximal possible wearing duration would vary significantly among the devices depending on the type of electrode. Spring-loaded or rigid pin electrodes were expected to apply more pressure on the head and thus to have a smaller comfort and a low possible wearing duration. Gel-based electrodes were expected to assure a better comfort and could be worn for longer. Furthermore, we were interested in testing the devices in regard to the visceral and behavioral levels of emotional design. These comprised the design of the devices and the ease of use. To this end, we formulated the following research questions for the evaluation of the devices:

Research question 1a: Does maximal possible wearing duration differ among devices with different electrode types?

Research question 1b: Does the visual perception of devices' design differ among each other?

Research question 1c: Does practicability of the devices differ among each other?

Especially in cases where the EEG device is worn in public (eg, workplace), some users could prefer a more unobtrusive design. This can be linked to the reflective level of Norman's 3-level model of emotional design [8]. Thereby, information from the visceral and behavioral levels are combined with our knowledge and experiences, filtered, and cognitively processed. At this level, user's self-image plays a crucial role. Beyond the intended use of the product, user preferences are based on who will see it and how these viewers will judge the user with it.

Hence, we were interested to find out if users were willing to accept less comfort for a more attractive headset design. On the basis of this consideration, we formulated our second research objective:

Research question 2: Does visual appearance affect the overall rating of the devices more than wearing comfort?

In the Methods section of our study, we introduce the EEG devices, material used, sample set, and procedure for conducting the experiments. The gained results are presented in the Results section and discussed in the following section. Thereby, we mention potential limitations to the study. Finally, the Conclusions subsection aims to highlight the main points of our study and draw general conclusions from the investigation.

Methods

Electroencephalography Systems

The investigation focused on 7 currently available mobile EEG devices. Table 2 shows the devices and summarizes their characteristics that are briefly described in the following.

NeuroSky's MindCap device is a 1-channel EEG system. It comes with a frontal electrode and an ear clip reference electrode. The use of conductive gel is not necessary, and the signal is transmitted wirelessly through Bluetooth interface. The weight is 119 g. The device is recommended for neurofeedback training and gaming.

Emotiv's EPOC device comes with 14 saline-based wet felt sensors. These are mounted on quite flexible plastic branches. The signal is transmitted wirelessly through Bluetooth interface. The EPOC device has a weight of 116 g.

Mindo's 4S Jellyfish device is a wireless dry electrode EEG device. The 4 electrodes that are mounted on a headband can be applied at either frontal or parietal sites. In our case of frontal EEG, foam-based electrodes (Figure 1, left) are recommended. In case of parietal EEG, spring-loaded pin electrodes (Figure 1, right) are to be applied. The reference is an adhesive electrode at the mastoid. The device weighs 95 g.

Mindo's 32 Trilobite device comprises 32 EEG channels. The frontal 3 of them are foam-based electrodes (Figure 1, left). The remaining 29 are spring-loaded pin electrodes (Figure 1, right). Furthermore, the device includes a ground and a reference electrode, both applied with a clip on the ear lobes. Signal transmission occurs wirelessly through Bluetooth. Its weight is 524 g.

BRI's BR8+ device has got 8 dry electrodes. The frontal 2 of them are foam-based electrodes (Figure 1, left). The remaining 6 are spring-loaded pin electrodes (Figure 1, right). The device includes ground and reference ear clip electrodes and a wireless signal transmission through Bluetooth. The earpads of the device do not have any technical functionality. They are thought to reduce the headset pressure and help positioning the headset at the center of the head. The BR8+ weighs 269 g.

g.tec's g.SAHARA/g.Nautilus device comprises 16 pin electrodes (Figure 2) that are mounted on a traditional EEG cap. The cap size can vary among small, medium, and large. However, to reduce financial costs, we used only the medium-sized cap. Adhesive ground and reference electrodes are applied at the mastoids. The signal is transmitted wirelessly by means of g.Nautilus device that is attached at the back of the EEG cap. It has a weight of 233 g.

g.tec's g.LADYbird/g.Nautilus device is a traditional gel-based EEG system with 16 active electrodes. An ear clip electrode serves as reference. Similar to the g.SAHARA/g.Nautilus device, the cap size can vary. However, in our study, we used only the medium-sized cap. The g.Nautilus device at the back of the cap allows for wireless signal transmission. The total weight of the EEG headset amounts to 165 g. Unlike the other devices, the g.LADYbird/g.Nautilus device is not designed for home and biofeedback applications. It is primarily developed for research and medical use and the treatment of locked-in patients. We included it to our study as state-of-the-art reference for EEG regarding user experience issues.

Finally, all manufacturers of our EEG devices promote their EEG systems as highly comfortable and easy to use.

Table 2. Electroencephalography (EEG) devices used.

EEG device	Headset	Electrode type	Number of electrodes	Weight
MindCap (NeuroSky Inc, San Jose, CA, USA)		Dry	1	119 g
EPOC (Emotiv Inc, San Francisco, CA, USA)		Saline-based	14	116 g
Jellyfish (Mindo, Hsinchu, Taiwan)		Foam-based	4	95 g
Trilobite (Mindo, Hsinchu, Taiwan)		3 foam-based, 29 spring-loaded pins	32	524 g
BR8+ (BRI Inc, Hsinchu, Taiwan)		2 foam-based, 6 spring-loaded pins	8	269 g
g.SAHARA (g.tec GmbH, Graz, Austria)		Pin electrodes	16	233 g
g.LADYbird (g.tec GmbH, Graz, Austria)		Gel-based	16	165 g

Figure 1. Foam-based frontal electrodes (left) and spring-loaded pin electrodes (right).

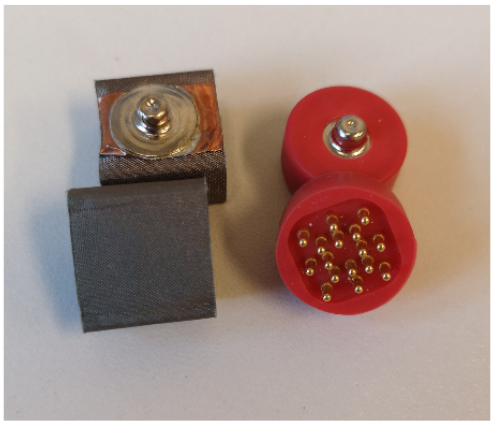


Figure 2. Pin electrodes of g.tec's g.SAHARA device.

Procedure and Subjects

Our study took place in a typical office setting. The 24 subjects participating (Table 3) completed over the course of 9 consecutive workdays a total of 9 sessions. The first session was aimed at familiarizing the subjects with the computer tasks and games they had to perform while wearing the EEG devices. In this session, we also assessed subjects' attitude toward technology by means of the 19 items of the TA-EG questionnaire (TA-EG: translated from the original German title: "Fragebogen zur Technikaffinität - Einstellung zu und Umgang mit elektronischen Geräten") [19-22]. The items are answered on a 5-point Likert scale (1=fully disagree and 5=fully agree) and address 4 dimensions: technology enthusiasm, competence in handling technology, positive attitude, and negative attitudes toward electronic devices. Subjects with calculated values below the median were assigned to the group of negative attitudes, whereas subjects with values over the median were assigned to the group of positive attitudes toward technology.

Table 3. Sample set used for analysis.

Age (years)	Male, n (%)	Female, n (%)	Total, N
26-34	2 (20)	8 (80)	10 (100)
35-49	3 (50)	3 (50)	6(100)
50-66	8 (100)	0 (0)	8(100)
Total	13	11	24

In the following 7 days, 1 device per day was selected in random order and tested independently of the others. Thereby, the subjects wore the device for 60 min and performed the same sequence of tasks and 1-min rest measurements with eyes closed and eyes opened. The devices were applied by an expert. At the end of each session, they were asked how long they would be able to wear the EEG headset. They indicated their answers on a 5-min steps scale between 0 and 120 min. They also answered questions regarding the device's design. Next, the subjects applied the device on their own. The expert inspected the signal quality of the EEG and gave instructions for improving it when needed. Moreover, 1-min rest measurements with eyes closed and eyes opened were performed, and thereafter, subjects rated the practicability of the device (Table 4). An exception was made for the g.LADYbird device that could not be taken off, reapplied, and properly used because of the smeared gel that builds conductive bridges. For the g.LADYbird device, we solely skipped the rest measurements.

During the last session, all EEG devices were rated. First, paired comparisons were conducted between every 21 pairs of 2 devices presented. Participants were asked to select the headset that they were willing to wear over a longer period of time or even daily. To avoid reliance on memory, subjects were instructed to reapply each of the 2 presented headsets and decide consciously. A mirror in front of them allowed them to include the visual appearance of the headset in their preference rating. Furthermore, we paid attention to the presentation order of the pairs and proceeded as recommended by Ross [23].

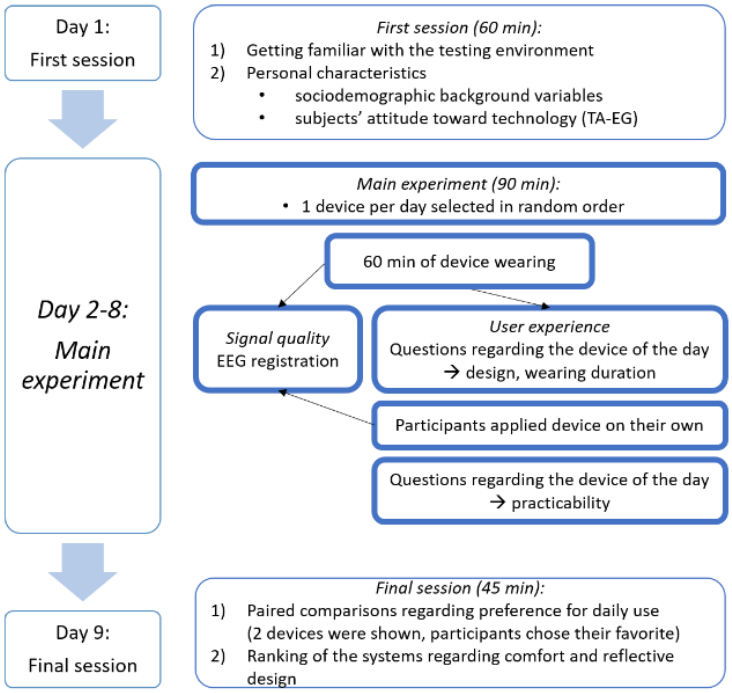
Finally, subjects completed a questionnaire where they had to rank the devices regarding wearing comfort and visual appearance separately (Table 4). Thereby, the item for visual appearance aimed to also integrate aspects from the reflective level of emotional design. Each of the headsets was set on a rank order between 1 (the most appropriate) and 7 (the least appropriate). Figure 3 outlines the experimental design of the study. All procedures were carried out with the adequate understanding and written consent of the subjects. The investigations acquired were approved by the local review board of our institution.

Table 4. User experience acquisition.

Aspects of emotional design	Item	Possible answers	Conducted	Research question
Visceral level	The headset has an attractive design	1: does not apply at all and 5: applies fully	After each session	1b
Behavioral level	I could apply and use the EEG[a] headset without aid	1: does not apply at all and 5: applies fully	After each session	1c
Behavioral level	How long are you able to wear EEG headset? Please mark the maximal-possible time duration in minutes on the scale below	Scale from 0 to 120 with 5 min steps	After each session	1a
Behavioral level	Wearing the device was comfortable	Ranking of the devices: 1: most appropriate and 7: least appropriate	Final session	2
Reflective level	It would not be a problem for me to be seen by my colleagues wearing the device	Ranking of the devices: 1: most appropriate and 7: least appropriate	Final session	2

[a]EEG: electroencephalography.

Figure 3. Experimental design of the study. EEG: electroencephalography.



Results

Comparisons Among Devices

The first research objective was concerned with the test of the devices regarding their wearing comfort after a longer period of time, visual appearance, and ease of use. For evaluation, we used subjects' answers conducted after each session (Table 4). Statistical analysis was conducted using nonparametric Friedman tests of differences among the repeated measures.

Maximal Possible Wearing Duration Differs Among Devices

Results indicated a highly significant change in maximal possible wearing duration among the EEG devices (χ^2_6=40.2, n=24; P<.001). Rankings are presented in Table 5. Dunn-Bonferroni posthoc tests were calculated for the examination of the differences among the devices (Table 5; see also Multimedia Appendix 1 for the exact values). Significant differences were obtained between the Trilobite device and all other devices except the BR8+. The Trilobite device was ranked lower regarding maximal wearing duration than the other devices.

Perception of Headset Design Differs Among Devices

Regarding the visual perception of devices' headset design, results indicated a significant change in subjects' ratings (χ^2_6=78.7, n=24; P<.001). Rankings are presented in Table 6. Dunn-Bonferroni posthoc tests were calculated for the examination of the differences among the devices (Table 6; Multimedia Appendix 2).

Table 5. Maximal possible wearing duration (min) for each device over all subjects.

EEG device	Mean (SD)	Median (min, max)
MindCap	92.29 (35.87)	112.5 (5, 120)
Jellyfish	86.66(31.78)	90.0 (30, 120)
BR8+	73.54 (30.16)	60.0 (30, 120)
EPOC	101.87 (25.10)	117.5 (30, 120)
g.Sahara	81.04 (33.45)	80.0 (10, 120)
Trilobite	48.75 (28.59)	50.0 (5, 120)
g.Ladybird	100.41 (23.99)	112.5 (45, 120)

Table 6. "The headset has an attractive design." (1: does not apply at all and 5: applies fully). Statistics calculated over all, male, and female subjects for each device.

EEG device	All		Male		Female	
	Mean (SD)	Median (min, max)	Mean (SD)	Median (min, max)	Mean (SD)	Median (min, max)
MindCap	3.71 (0.95)	4.0 (1, 5)	4.15 (0.68)	4.0 (3, 5)	3.18 (0.98)	3.0 (1, 4)
Jellyfish	3.58 (0.97)	4.0 (2, 5)	3.85 (0.80)	4.0 (2, 5)	3.27 (1.10)	4.0 (2, 5)
BR8+	3.58 (0.97)	4.0 (1, 5)	3.92 (0.64)	4.0 (3, 5)	3.18 (1.16)	3.0 (1, 5)
EPOC	4.08 (0.77)	4.0 (2, 5)	4.23 (0.92)	4.0 (2, 5)	3.91 (0.53)	4.0 (3, 5)
g.Sahara	2.21 (1.10)	2.0 (1, 5)	2.62 (1.12)	3.0 (1, 5)	1.73 (0.90)	2.0 (1, 4)
Trilobite	2.58 (0.92)	2.5 (1, 5)	2.92 (0.86)	3.0 (2, 5)	2.18 (0.87)	2.0 (1, 4)
g.Ladybird	2.08 (0.83)	2.0 (1, 4)	2.46 (0.66)	2.0 (2, 4)	1.64 (0.80)	1.0 (1, 3)

Significant differences were obtained between the g.LADYbird device and all other devices except g.SAHARA and Trilobite. The g.SAHARA device showed significant differences to all devices except Trilobite and g.LADYbird. The Trilobite device showed significant differences to the EPOC, MindCap, and Jellyfish devices. At this point, we also looked at possible gender effects relating to the perception of headsets' design. We evaluated the ratings separately for male and female participants (Table 6) and found highly significant differences among devices for both groups (male: χ^2_6=41.9, n=13, P<.001; female: χ^2_6=38.3, n=11, P<.001). Dunn-Bonferroni posthoc tests for male participants' ratings indicated significant differences between the Trilobite and EPOC devices as well as between g.SAHARA and MindCap and g.SAHARA and EPOC (Table 6; Multimedia Appendix 3). Furthermore, there were significant differences between the g.LADYbird device and all other devices except g.SAHARA and Trilobite. Dunn-Bonferroni posthoc tests for female participants' ratings indicated significant differences between the Trilobite and EPOC devices, g.SAHARA and EPOC as well as between g.LADYbird and EPOC and g.LADYbird and Jellyfish (Table 6; Multimedia Appendix 4).

Practicability Differs Among Devices

Results of subjects' ratings regarding the practicability of the devices indicated highly significant differences among the EEG devices (χ^2_6=83.2, n=24; P<.001). Rankings are presented in Table 7. Dunn-Bonferroni posthoc tests were calculated for the examination of the differences among the devices (Table 7; Multimedia Appendix 5).

Significant differences were obtained between the g.LADYbird device and all remaining devices. To evaluate possible differences among subjects related to their attitude toward technology, we used the results from the TA-EG questionnaire and clustered our subjects in 2 groups. Subjects with a value below the overall median of 69.5 (range between 41 and 81) were assigned to the group with a negative attitude toward technology (mean age of cluster: 41 years, 5 females, and 7 males) and subjects with a value over the median to the group with a positive attitude (mean age of cluster: 44 years, 6 females, and 6 males). We evaluated the practicability ratings separately and found highly significant differences among devices for both groups (negative attitude: χ^2_6=48.5, n=12, P<.001; positive attitude: χ^2_6=40.6, n=12, P<.001).

Dunn-Bonferroni posthoc tests for the ratings of subjects with a negative attitude toward technology indicated significant differences between the g.LADYbird and all remaining devices (Table 7; Multimedia Appendix 6). Dunn-Bonferroni posthoc tests for the ratings of subjects with a positive attitude toward technology indicated significant differences between the g.LADYbird and all other devices except the Trilobite and g.SAHARA (Table 7; Multimedia Appendix 7).

The critical reader could argue that for evaluating the practicability, the signal quality of the device had to be taken into account after self-fitting the device. For the sake of completeness, we compared the signal quality of the rest measurements from self-fitting versus expert fitting of the system. The evaluation of the electroencephalogram was done in the time domain manually. A medical technical assistant with specialization in EEG and years of experience visually inspected the electroencephalograms and manually marked artifact segments. We computed the percentage of denoted artifacts compared with the entire recording time for each channel. We calculated the means over the channels for each subject and device. For comparison between the signal qualities from self-fitting versus expert fitting, we conducted a Wilcoxon paired difference test for each EEG system. The results are presented in Table 8. Rest measurements with closed eyes did not show significant differences between the fittings for none of the devices. Rest measurements with eyes opened indicated significant differences between the fittings for the BR8+ and the g.SAHARA devices (BR8+: z=−3.886, P<.001, r=0.56; g.SAHARA: z=4:086, P<.001, r=0.59).

For readers more interested in the signal quality evaluation of the devices, we would like to draw their attention on our paper on that topic [24].

Table 7. "I could apply and use the EEG headset without aid." (1: does not apply at all and 5: applies fully). Statistics calculated over all subjects, subjects with positive attitude, and subjects with negative attitude toward technology for each device. EEG: electroencephalography.

EEG device	All		Positive attitude		Negative attitude	
	Mean (SD)	Median (min, max)	Mean (SD)	Median (min, max)	Mean (SD)	Median (min, max)
MindCap	4.63 (0.64)	5.0 (3, 5)	4.42 (0.79)	5.0 (3, 5)	4.42 (0.79)	5.0 (3, 5)
Jellyfish	4.67 (0.56)	5.0 (3, 5)	4.50 (0.67)	5.0 (3, 5)	4.50 (0.67)	5.0 (3, 5)
BR8+	4.21 (1.10)	5.0 (2, 5)	3.83 (1.19)	4.0 (2, 5)	3.83 (1.19)	4.0 (2, 5)
EPOC	4.54 (0.58)	5.0 (3, 5)	4.50 (0.52)	4.5 (4, 5)	4.50 (0.52)	4.5 (4, 5)
g.Sahara	4.04 (1.04)	4.0 (2, 5)	3.58 (1.24)	4.0 (2, 5)	3.58 (1.24)	4.0 (2, 5)
Trilobite	3.54 (1.31)	4.0 (1, 5)	3.00 (1.27)	3.0 (1, 5)	3.00 (1.27)	3.0 (1, 5)
g.Ladybird	1.75 (0.89)	1.5 (1, 5)	1.75 (0.86)	1.5 (1, 3)	1.75 (0.86)	1.5 (1, 3)

Table 8. Artifact proportions (%) of rest measurements with eyes open and closed from self-fitting and expert fitting of the system averaged over channels and subjects and considered for each device separately.

EEG device	Eyes closed				Eyes open			
	Expert fitting		Self-fitting		Expert fitting		Self-fitting	
	Mean (SD)	Median (min, max)	Mean (SD)	Median (min, max)	Mean (SD)	Median (min, max)	Mean (SD)	Median (min, max)
MindCap	15.37 (33.54)	0.0 (0.0, 99.9)	17.60 (35.92)	0.0 (0.0, 99.9)	16.75 (33.38)	0.0 (0.0, 99.9)	10.77 (23.68)	0.0 (0.0, 99.9)
Jellyfish	23.26 (27.56)	13.4 (.0, 99.7)	14.64 (18.16)	6.9 (0.0, 61.9)	24.15 (25.98)	14.7 (0.0, 87.3)	20.94 (22.88)	11.8 (0.0, 80.0)
BR8+	48.38 (21.55)	49.9 (3.7, 87.5)	59.51 (22.44)	63.23 (12.5, 99.9)	45.12 (17.36)	47.7 (14.3, 80.0)	75.62 (20.89)	78.4 (26.5, 100)
EPOC	23.25 (37.07)	3.6 (0.0, 99.9)	37.82 (45.16)	11.4 (0.0, 99.9)	22.18 (36.61)	5.0 (0.0, 99.9)	37.60 (42.69)	13.3 (0.0, 99.9)
g.Sahara	32.05 (11.47)	34.1 (5.7, 55.5)	32.54 (13.46)	33.0 (0.0, 65.1)	9.79 (12.74)	4.0 (0.0, 41.8)	21.33 (16.85)	18.5 (3.5, 74.5)
Trilobite	29.14 (26.33)	18.6 (3.1, 106.25)	22.10 (26.19)	14.4 (0.0, 106.1)	33.83 (25.32)	23.9 (0.0, 91.4)	23.69 (20.56)	16.6 (0.0, 83.5)

Wearing Comfort and Visual Appearance

Our research question 2 asked if visual appearance affects the overall rating of the devices more than their wearing comfort. For the evaluation, we used multiple linear regression analysis. Ranking values of the items for visual appearance and wearing comfort (Table 4) served as independent variables. The criterion was the devices' ranking order regarding preference for daily use. This was calculated from the conducted paired comparisons.

For the sake of completeness, we have to mention that results from paired comparisons were not transitive for 6 subjects. In these cases, some devices have been selected with the same frequency, and thus, subjects' preference could not be mapped on an ordinal scale. Analysis of these subjects' decisions regarding the less rejected devices did not yield to a result, either. Hence, the 6 subjects with inconsistent answers were disclosed from further analysis.

We computed a multiple linear regression for each device separately. The results are presented in Table 9. Wearing comfort and visual appearance of the devices were able to statistically significant predict subjects' preference for daily use, except for the g.LADYbird device ($F_{2,15}$=0.752; P=.49). Wearing comfort had a large impact on device preference for almost all devices, whereas visual appearance was a poor predictor. An exception was the EPOC device. Hereby, visual appearance had a large impact on the preference, whereas wearing comfort had none. For the BR8+ device, both predictors were important. However, the wearing comfort was more influential.

At this point, we also looked at possible gender effects relating to the utilitarian versus hedonic aspects of the experience. For the male participants, wearing comfort and visual appearance were able to statistically significant predict subjects' preference for daily use, except for the g.LADYbird device ($F_{2,7}$=0.147; P=.87). Wearing comfort had a large impact on device preference for all devices except for the EPOC device where visual appearance was a better predictor. For the female participants, a significant regression equation with significant predictors was found for the Jellyfish ($F_{2,5}$=29.837; P=.002) and EPOC ($F_{2,5}$=25.571, P=.002) devices. For Jellyfish, wearing comfort significantly predicted subjects' preference, whereas for EPOC, visual appearance had a greater impact on subjects' preference ratings. Overall, it can be said that in cases where the regression models became significant, we were not able to identify opposing effects between female and male participants (Table 9).

Table 9. Results of multiple linear regression analysis for each device.

EEG[a] device and gender	R^2	Model		Wearing comfort		Visual appearance	
		F test (*df*)	*P* value	Coefficient	*P* value	Coefficient	*P* value
MindCap							
Both	0.938	112.518 (2,15)	<.001	1.112	<.001	−0.245	.21
Male	0.989	305.051 (2,7)	<.001	0.985	<.001	−0.024	.90
Female	0.707	6.018 (2,5)	.05	0.472	.16	0.067	.79
Jellyfish							
Both	0.825	35.319 (2,15)	<.001	0.797	<.001	0.069	.77
Male	0.764	11.357 (2,7)	.006	0.751	.01	0.210	.60
Female	0.923	29.837 (2,5)	.002	0.731	.008	0.235	.59
BR8+							
Both	0.846	41.182 (2,15)	<.001	0.701	<.001	0.327	.04
Male	0.952	70.150 (2,7)	<.001	0.802	.001	0.312	.07
Female	0.498	2.479 (2,5)	.18	0.540	.09	−0.010	.98
EPOC							
Both	0.849	42.080 (2,15)	<.001	0.149	.14	0.656	<.001
Male	0.823	16.286 (2,7)	.002	0.191	.53	0.627	.02
Female	0.911	25.571 (2,5)	.002	0.153	.14	0.655	.009
g.SAHARA							
Both	0.742	21.603 (2,15)	<.001	0.740	<.001	−0.040	.76
Male	0.939	54.275 (2,7)	<.001	0.777	<.001	0.011	.89
Female	0.633	4.312 (2,5)	.08	0.677	.03	0.000	>.99
Trilobite							
Both	0.737	21.026 (2,15)	<.001	0.943	<.001	0.139	.39
Male	0.770	11.706 (2,7)	.006	1.043	.002	0.109	.63
Female	0.485	2.354 (2,5)	.19	0.620	.12	0.260	.41
g.LADYbird							
Both	0.091	0.752 (2,15)	.49	0.018	.93	0.243	.25
Male	0.040	0.147 (2,7)	.87	0.063	.68	0.038	.75
Female	0.335	1.261 (2,5)	.36	−0.400	.42	2.300	.20

[a]EEG: electroencephalography.

Discussion

Comparisons Among Devices

In our first research objective, we were concerned to test the devices regarding 3 user experience aspects: wearing comfort, visual appearance, and ease of use.

Pin Electrodes Had the Lowest Wearing Comfort

Evaluation of the maximal possible wearing time as an indicator of devices' wearing comfort revealed the Trilobite device to be significantly less pleasant to wear than the remaining. The reason could be the uncomfortable pin electrodes. Overall means of maximal possible wearing duration indicated devices without pin electrodes such as the EPOC, MindCap, and g.LADYbird as the most favorable for a longer wearing time and with significant differences to the Trilobite. The finding that pin electrodes were less preferred was similar to findings by Grozea et al [14] but inconsistent to the results by Nijboer et al [18] and Izdebski et al [16]. However, Hairston et al [17] also emphasized the importance of the headset's ability to adjust to the different heads to assure comfort. In their work, they highlighted the need of flexible headsets to assure comfort during wearing. This aspect was also prominent in the work of Izdebski et al [16] who found that cap fit was rated as poor for headsets with rigid headsets. In our study, Trilobite's headset was the most rigid one. Furthermore, the Trilobite device was much heavier than the other devices. These 2 facts could have multiplied the impact of the pin electrodes on wearing comfort. The BR8+ device had pin electrodes, a rather rigid headset but less weight. Similar to the Trilobite, it yielded small values

regarding the maximal possible wearing duration. The g.SAHARA with pin electrodes but flexible headset and less weight had small wearing duration ratings, but these were higher than those of the Trilobite and BR8+ devices. We concluded that pin electrodes had the lowest wearing comfort, in particular when coupled with a rigid, heavy headset.

An Unobtrusive Design Coped Better With Individual Preferences

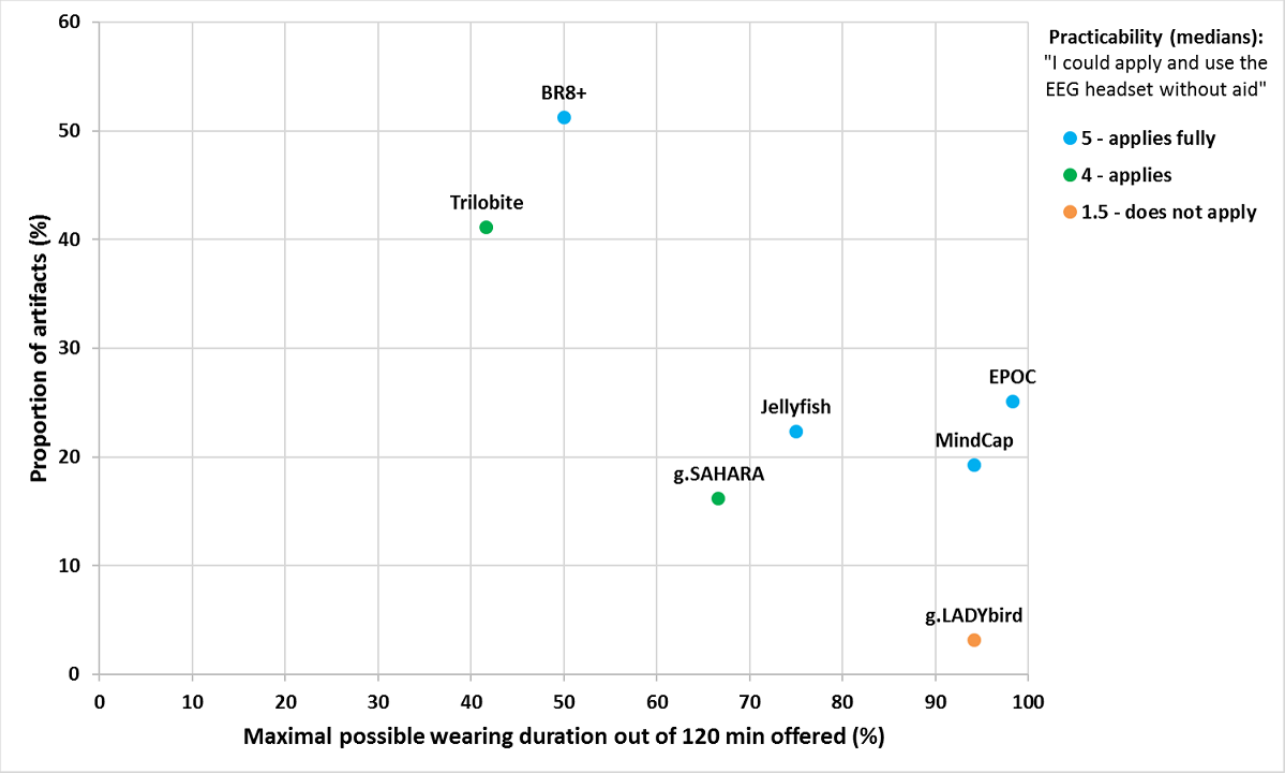
Headset design is not only responsible for the wearing comfort but also primarily responsible for device's visual appearance. Overall ratings of headset design indicated that the devices with a traditional EEG cap (ie, g.LADYbird and g.SAHARA) were significantly less preferred than all others, except the Trilobite device. The latter was also significantly less preferred than the MindCap, Jellyfish, and EPOC devices. Females' ratings indicated more variability than males' ratings leading to less significant differences among the devices. However, both genders perceived the design of g.LADYbird's and g.SAHARA's traditional caps and Trilobite's helmet as less attractive. Both groups primarily preferred the headsets of EPOC and Jellyfish with EPOC, indicating more significant differences to the other devices, in particular, by female subjects. This result was consistent with the results by Nijboer et al [18] where participants rated their appearance with the EPOC as best. Nijboer et al stated that reasons for the refusal of caps were that the whole head and part of the face were covered, and hair was flattened and invisible. In our study, the g.LADYbird, g.SAHARA, Trilobite, and MindCap devices covered subjects' whole head. However, ratings of the MindCap were significantly better compared with the other 3 devices. This was particularly true among the male subjects. We assumed that rating of the design was related to aspects of aesthetics, fashion style, and individual preference. These aspects might be strongly connected to the reflective level of emotional design. An unobtrusive headset design could have more potential to cope with different individual preferences because it is not eye-catching.

Practicability Was Closely Linked to Gel Electrodes and Attitude Toward Technology

Finally, we asked the subjects to rate the ease of use of the devices. Results indicated significant differences between the gel-based g.LADYbird and all remaining devices. This was reasonable, especially when considering that a second person was needed for applying the gel. Furthermore, subjects had to wash their hair after they took off the cap. We concluded that the effort for use was definitely high. The g.SAHARA and Trilobite devices were also rated as less easy to use. We supposed that this might be because of their larger number of electrodes but have to be aware that g.SAHARA had only 2 electrodes more than the EPOC device. Subjects with a negative attitude toward technology showed similar results regarding the practicability of the devices. However, subjects with a positive attitude toward technology did not indicate significant differences between the gel-based g.LADYbird and pin-based g.SAHARA neither between the g.LADYbird and Trilobite devices. Although these findings were surprising, we supposed that technical affine subjects were more critical during their ratings, and this could lead to more variability in their ratings. Taken the results of the signal quality comparison (Figure 4) into account, we noted similar tendencies between practicability ratings from subjects with a positive attitude toward technology and increased proportion of artifacts by self-fitting the devices. This was particularly true during the rest measurements with eyes opened, as subjects might have behaved more actively than with eyes closed. Thereby, the BR8+, g.SAHARA, and, to a lesser extent, the EPOC devices yielded more artifacts when compared with the fittings by an expert and revealed less practicability when rated by technical affine subjects. Nevertheless, the g.LADYbird device had the worst practicability ratings across subjects although a limitation of our study might be that we did not give the opportunity to the subjects to apply the device and the gel on their own. We believe that self-fitting of the gel-based electrodes would not have altered the ratings but must admit that future user experience research should consider this issue. Finally, we argue that subjects with a positive attitude toward technology were more accurate in their rating of device practicability.

In conclusion, although the practicability of the devices was closely linked to gel or dry electrodes, wearing comfort and design of the devices seemed to be more expressive. Thereby, we observed that devices that could be worn for a longer period of time did not always have an attractive design.

Figure 4. Relation between signal quality and comfort and trade-off against practicability. EEG: electroencephalography.



Wearing Comfort and Visual Appearance

In our second research question, we were interested to find out if wearing comfort was more important to the user than the visual appearance of the device. Thus, we asked subjects to rank all devices regarding both aspects separately. Furthermore, paired comparisons of the devices led us to a rank order regarding preference for daily use.

Results of a multiple linear regression analysis for each device indicated that, in general, wearing comfort was the better predictor for users' device preference. Exceptions were the EPOC and the g.LADYbird devices. Although for the g.LADYbird, none of the 2 aspects seemed to have any impact on device's preference ranking; the results for the EPOC device revealed an opposite tendency, that is, EPOC's visual appearance influenced subject's decision more than its wearing comfort. A reason for this could be that EPOC's wearing comfort was unobtrusive although its design was futuristic and professional. We assumed that this attracted the subjects and gave more weight to the visual appearance when it came to a preference for daily use. Interestingly, the design of the BR8+ was also one of the most modern and futuristic ones. The fact that BR8+'s visual appearance was a supplementary predictor to its comfort seemed to confirm our assumption.

Regarding the results of the g.LADYbird device, we had to speculate. The device was assumed to not cause any head pressure; hence, wearing comfort should be unobtrusive and a weak predictor for the preference for daily use. Its visual appearance was indeed not very attractive for daily wearing. However, this fact did not have a large influence on the preference either, similar to the g.SAHARA device that had the same cap. The main difference to all other devices was the application of gel and the necessity to wash the hair after each use of the device. Although comfortable to wear, the gel-based electrodes were undoubtedly inconvenient for daily use outside the laboratory. Hence, the ease of use could have affected the preference more than the examined factors.

Male and female participants did not show opposing results related to the predictors of daily use preference. Although almost all models (except g.LADYbird) became significant for the male participants, for the female participants, only 2 models reached the significance level (Jellyfish and EPOC). An explanation could be that females' ratings were not as consistent as males' ratings among each other. However, we have to be also aware of the small number of participants (8 females vs 10 males) that could have led to this result. To explore gender differences related to utilitarian versus hedonic aspects of experience, more research with larger subsets is needed. We have to draw attention to our sample's structure (Table 3) consisting of young female and older male participants. Disentangle the gender and age factors at these numbers seemed not possible. We assumed that regarding emotional design, the gender factor is more influential than the age, but the reader should note that the latter could have an effect, too. Further research should emphasize on this issue.

In general, the results of both genders emphasized that visual appearance was a better predictor only for the EPOC device. By taking into account the reflective level of emotional design, we add new insight about how the factors of comfort and visual appearance translate to user preference. Our results broaden the assumption by Nijboer et al [18] who postulated that the preference of EPOC was an evidence for the fact that aesthetics might be more important than comfort.

Conclusions

In our study, we investigated the user experience of mobile EEG devices. We compared 7 different EEG devices and offered a differentiated look at emerging mobile and gel-free EEG technology. The results yielded are summarized in Table 10. For the sake of convenience, we report only the artifact proportion differences between self-fitting and expert fitting from the eyes-closed measurement.

In addition, we gave insight into the relation between user experience aspects and device preference. The wearing comfort given by a device was the main factor for its daily use. The visual appearance of the device was certainly an important point. However, it only became influential when comfort was assured. Users were not willing to accept less comfort for a more attractive headset design. The reflective level of emotional design became important only if the behavioral level of the product was satisfactory.

To provide practical information to users of EEG devices, we combined the signal quality results from the study by Radüntz [24] with the current user experience results and concluded which system could be used under which condition. The EPOC device achieved the best results regarding user experience, but it suffered from a large proportion of artifacts. Although the EPOC device can be used in public because of its attractive design and the feeling of ease of use, potential users should be aware of the issues regarding signal quality, in particular, if the device is self-applied by a layman. Outstanding performances regarding maximal possible wearing duration and signal quality were obtained for the traditional gel-based but mobile g.LADYbird device. This device can be recommended for neuroscience research where precise and prolonged measurements are required without any deductions in comfort. However, devices wearing in public and self-application are not recommended. The MindCap device reviled good user experience results and satisfying signal quality. Users must consider that scientifically valid assertions could be hampered because of only 1 electrode available. The Jellyfish and g.SAHARA devices yielded similar results regarding comfort but differences regarding design (ie, better results for Jellyfish) and signal quality (ie, better results for g.SAHARA). We believe that g.SAHARA is a good solution for field experiments, where subjects are not exposed to the general public, and signal quality is important. Nevertheless, researchers should be aware of potential comfort issues that could arise in the course of time because of the pin electrodes. Potential applications for the Jellyfish device might be better suited for the gaming or biofeedback sector. The BR8+ and Trilobite devices did not meet our requirement for user experience, in particular, because of comfort issues. Furthermore, signal quality was lacking. Figure 4 illustrates the trade-offs between signal quality and user experience so that readers might be able to see if there are any devices of sufficient quality that might also be acceptable for daily use. The x-axis depicts devices' comfort rankings, calculated as a percentage of the maximal possible wearing duration in minutes out of 120 min offered. The y-axis represents the proportion of artifacts taken from the study by Radüntz [24].

Finally, we have to admit that there might be further factors that could have contributed to the preference decision. Our research could be seen as a precondition for the use of emerging EEG technology under realistic conditions in field experiments with longer duration. It paves the way for the development of usable applications with wearables and contributes to consumer health informatics and health-enabling technologies. Furthermore, our results provided guidance for the technological development direction of new EEG devices related to aspects of emotional design.

It has to be mentioned that the EEG equipment market shows rapid development. During this study, new devices appeared on the market that could not be tested, for example, the actiCAP Xpress Twist/LiveAmp device by Brain Products or the highly innovative approach using in-ear EEG technology [25,26]. However, our study design could easily be used in subsequent studies of new devices and benchmark the evaluation of further emerging EEG technology. Integration of test results from new devices into the findings already in existence would make it possible to compare the user experience of emerging EEG technology.

Table 10. User experience results of tested electroencephalography devices (medians over all subjects).

EEG[a] device	Comfort: maximal wearing duration (min)	Design (higher values indicate a more attractive design)	Practicability (higher values indicate greater practicability)	Artifact proportions (eyes closed: self-fitting-expert fitting [%]; higher values indicate more artifacts when self-fitted)
MindCap	113	4	5	2.2
Jellyfish	90	4	5	−8.6
BR8+	60	4	5	11.1
EPOC	118	4	5	14.6
g.SAHARA	80	2	4	0.5
Trilobite	50	2.5	4	−7
g.LADYbird	113	2	1.5	Not applicable

[a]EEG: electroencephalography.

Acknowledgments

The authors would like to thank Gerd Menzel and Ralph Blüthner for their technical and overall support, our student assistant Friederice Schröder for conducting the experiments, Marion Freyer for the visual inspection of the data and manual artifact marking, and the student assistants Yuexin Cao and Ilona Pritschke for computational support and graphic editing. Furthermore, the authors would like to thank Gabriele Freude and Uwe Rose for their general project support. Finally, they want to thank the reviewers of the manuscript for their insightful comments, which they deeply appreciate. The study was funded by the Federal Institute for Occupational Safety and Health (project number: F 2402).

Authors' Contributions

TR initiated the project and was responsible for the overall conception of the investigation. Data analysis was performed by TR. Data interpretation was performed by TR and BM. The manuscript was written by TR. Final critical editing was performed by BM.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Maximal possible wearing duration for each device over all subjects.
[PDF File (Adobe PDF File), 299KB-Multimedia Appendix 1]

Multimedia Appendix 2

Attractive design ratings for each device over all subjects.
[PDF File (Adobe PDF File), 434KB-Multimedia Appendix 2]

Multimedia Appendix 3

Attractive design ratings for each device over the male subjects.
[PDF File (Adobe PDF File), 436KB-Multimedia Appendix 3]

Multimedia Appendix 4

Attractive design ratings for each device over the female subjects.
[PDF File (Adobe PDF File), 433KB-Multimedia Appendix 4]

Multimedia Appendix 5

Practicability ratings for each device over all subjects.
[PDF File (Adobe PDF File), 436KB-Multimedia Appendix 5]

Multimedia Appendix 6

Practicability ratings for each device over subjects with negative attitude toward technology.
[PDF File (Adobe PDF File), 436KB-Multimedia Appendix 6]

Multimedia Appendix 7

Practicability ratings for each device over subjects with positive attitude toward technology.
[PDF File (Adobe PDF File), 436KB-Multimedia Appendix 7]

References

1. Rezeika A, Benda M, Stawicki P, Gembler F, Saboor A, Volosyak I. Brain-computer interface spellers: a review. Brain Sci 2018 Mar 30;8(4):pii: E57 [FREE Full text] [doi: 10.3390/brainsci8040057] [Medline: 29601538]
2. Ramadan RA, Vasilakos AV. Brain computer interface: control signals review. Neurocomputing 2017 Feb;223:26-44. [doi: 10.1016/j.neucom.2016.10.024]
3. Minguillon J, Lopez-Gordo MA, Pelayo F. Trends in EEG-BCI for daily-life: requirements for artifact removal. Biomed Signal Process Control 2017 Jan;31:407-418. [doi: 10.1016/j.bspc.2016.09.005]

4. Vaid S, Singh P, Kaur C. EEG Signal Analysis for BCI Interface: A Review. In: Proceedings of the Fifth International Conference on Advanced Computing & Communication Technologies. 2015 Presented at: ICACCS'15; February 21-22, 2015; Haryana, India p. 143-147. [doi: 10.1109/ACCT.2015.72]
5. Prashant P, Joshi A, Gandhi V. Brain Computer Interface: A Review. In: Proceedings of the 5th Nirma University International Conference on Engineering. 2016 Presented at: NUiCONE'16; November 26-28, 2016; Ahmedabad, India p. 1-6. [doi: 10.1109/NUICONE.2015.7449615]
6. Abdulkader SN, Atia A, Mostafa MS. Brain computer interfacing: applications and challenges. Egypt Inform J 2015 Jul;16(2):213-230. [doi: 10.1016/j.eij.2015.06.002]
7. Lee S, Shin Y, Woo S, Kim K, Lee HN. Review of wireless brain-computer interface systems. In: Fazel-Rezai R, editor. Brain-Computer Interface Systems: Recent Progress and Future Prospects. Rijeka, Croatia: IntechOpen; 2013.
8. Norman DA. Emotional Design: Why We Love (or Hate) Everyday Things. New York, US: Basic Books; 2004.
9. Ortony A, Norman DA, Revelle W. The role of affect and proto-affect in effective functioning. In: Fellous JM, Arbib MA, editors. Who Needs Emotions?: The Brain Meets the Robot. New York, US: Oxford University Press; 2005:173-202.
10. Ries AJ, Touryan J, Vettel J, McDowell K, Hairston WD. A comparison of electroencephalography signals acquired from conventional and mobile systems. J Neurosci Neuroeng 2014 Feb 1;3(1):10-20. [doi: 10.1166/jnsne.2014.1092]
11. Grummett TS, Leibbrandt RE, Lewis TW, DeLosAngeles D, Powers DM, Willoughby JO, et al. Measurement of neural signals from inexpensive, wireless and dry EEG systems. Physiol Meas 2015 Jul;36(7):1469-1484. [doi: 10.1088/0967-3334/36/7/1469] [Medline: 26020164]
12. Debener S, Minow F, Emkes R, Gandras K, de Vos M. How about taking a low-cost, small, and wireless EEG for a walk? Psychophysiology 2012 Nov;49(11):1617-1621. [doi: 10.1111/j.1469-8986.2012.01471.x] [Medline: 23013047]
13. Nikulin VV, Kegeles J, Curio G. Miniaturized electroencephalographic scalp electrode for optimal wearing comfort. Clin Neurophysiol 2010 Jul;121(7):1007-1014. [doi: 10.1016/j.clinph.2010.02.008] [Medline: 20227914]
14. Grozea C, Voinescu CD, Fazli S. Bristle-sensors--low-cost flexible passive dry EEG electrodes for neurofeedback and BCI applications. J Neural Eng 2011 Apr;8(2):025008. [doi: 10.1088/1741-2560/8/2/025008] [Medline: 21436526]
15. Ekandem JI, Davis TA, Alvarez I, James MT, Gilbert JE. Evaluating the ergonomics of BCI devices for research and experimentation. Ergonomics 2012;55(5):592-598. [doi: 10.1080/00140139.2012.662527] [Medline: 22506831]
16. Izdebski K, Oliveira AS, Schlink BR, Legkov P, Kärcher S, Hairston WD, et al. Usability of EEG Systems: User Experience Study. In: Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments. 2016 Presented at: PETRA'16; June 29-July 1, 2016; Corfu, Island, Greece URL:https://dl.acm.org/citation.cfm?id=2910714 [doi: 10.1145/2910674.2910714]
17. Hairston WD, Whitaker KW, Ries AJ, Vettel JM, Bradford JC, Kerick SE, et al. Usability of four commercially-oriented EEG systems. J Neural Eng 2014 Aug;11(4):046018. [doi: 10.1088/1741-2560/11/4/046018] [Medline: 24980915]
18. Nijboer F, van de Laar B, Gerritsen S, Nijholt A, Poel M. Usability of three electroencephalogram headsets for brain–computer interfaces: a within subject comparison. Interact Comput 2015 Jul 15;27(5):500-511. [doi: 10.1093/iwc/iwv023]
19. Karrer K, Glaser C, Clemens C, Bruder C. Technikaffinität erfassen – der fragebogen TA-EG. In: Lichtenstein A, Stößel C, Clemens C, editors. Der Mensch im Mittelpunkt Technischer Systeme. 8. Berliner Werkstatt Mensch-Maschine-Systeme. Düsseldorf, Germany: VDI Verlag GmbH; 2009:196-201.
20. Schüssel F, Honold F, Weber M. Influencing factors on multimodal interaction during selection tasks. J Multimodal User In 2012 Nov 25;7(4):299-310. [doi: 10.1007/s12193-012-0117-5]
21. Mertens A, Wille M, Theis S, Rasche P, Finken L, Schlick CM. Attitudes of Elderly People Towards Assistive System: Influence of Amortization Barriers on the Adherence in Technically Assisted Rehabilitation and the Diffusion of Health Technologies. In: Proceedings of the Triennial Congress of the International Ergonomics Association. 2015 Presented at: IEA'15; August 9-14, 2015; Melbourne, Australia URL:https://tinyurl.com/y3rn6eqm
22. Nitsch V, Glassen T. Investigating the Effects of Robot Behavior and Attitude Towards Technology on Social Human-Robot Interactions. In: Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication. 2015 Presented at: RO-MAN'15; August 31-September 4, 2015; Kobe, Japan p. 535-540. [doi: 10.1109/ROMAN.2015.7333560]
23. Ross RT. Optimum orders for the presentation of pairs in the method of paired comparisons. J Educ Psychol 1934;25(5):375-382. [doi: 10.1037/h0070754]
24. Radüntz T. Signal quality evaluation of emerging EEG devices. Front Physiol 2018;9:98 [FREE Full text] [doi: 10.3389/fphys.2018.00098] [Medline: 29491841]
25. Goverdovsky V, von Rosenberg W, Nakamura T, Looney D, Sharp DJ, Papavassiliou C, et al. Hearables: multimodal physiological in-ear sensing. Sci Rep 2017 Jul 31;7(1):6948 [FREE Full text] [doi: 10.1038/s41598-017-06925-2] [Medline: 28761162]
26. Looney D, Kidmose P, Park C, Ungstrup M, Rank ML, Rosenkranz K, et al. The in-the-ear recording concept: user-centered and wearable brain monitoring. IEEE Pulse 2012;3(6):32-42. [doi: 10.1109/MPUL.2012.2216717] [Medline: 23247157]

Abbreviation

EEG: Electroencephalography

Edited by G Eysenbach; submitted 24.04.19; peer-reviewed by J Cañas, R Agrawal, A Mahnke, J Salisbury, T Steffert, J Cahill, T Gaber, T Mühlhausen; comments to author 19.06.19; revised version received 28.06.19; accepted 06.07.19; published 23.08.19

Please cite as:
Radüntz T, Meffert B
User Experience of 7 Mobile Electroencephalography Devices: Comparative Study
JMIR Mhealth Uhealth 2019;7(8):e14474
URL: http://mhealth.jmir.org/2019/8/e14474/
doi: 10.2196/14474
PMID:

The Effect of an Exceptional Event on the Subjectively Experienced Workload of Air Traffic Controllers

Thea Radüntz[1(✉)], Norbert Fürstenau[2], André Tews[2], Lea Rabe[1], and Beate Meffert[3]

[1] Mental Health and Cognitive Capacity, Federal Institute for Occupational Safety and Health, Berlin, Germany
raduentz.thea@baua.bund.de
[2] German Aerospace Center, Institute of Flight Guidance, Braunschweig, Germany
[3] Signal Processing and Pattern Recognition, Department of Computer Science, Humboldt-Universität zu Berlin, Berlin, Germany

Abstract. There is a growing consensus concerning the negative consequences of inappropriate workload on employee's health and the safety of persons. In a simulator study, we focused on air traffic controllers during arrival management tasks. Our aim was to find out if the number of aircraft or the occurrence of an exceptional event added load to the subjectively experienced workload. The workload was assessed using the NASA-TLX, instantaneous self-assessment (ISA) questionnaire, and expert ratings. Our sample consisted of 21 subjects. According to standard ANOVA procedures, controllers' subjective ratings showed a high-significant discrimination between the different air traffic demands but only a weak-significant discrimination between sessions with and without event. In particular, we were not able to obtain a significant interaction effect between traffic volume and event. However, the examination of between-subject factors could reveal additional information about controller's rating behavior. We currently conclude that while the effect of the number of aircraft was evident, the impact of an exceptional event remained doubtful.

Keywords: Mental workload · Air traffic controllers · Subjective ratings · NASA-TLX · ISA

1 Introduction

There is a growing consensus concerning the negative consequences of inappropriate workload that can affect the individual itself but also other people that count on it. High mental workload is associated with increased anxiety, stress, and a lack of detachment from work during off-job time [8,9,24]. The missing recovery from work-related stress can then lead to weakness, tiredness, and exhaustion.

L. Longo and M. C. Leva (Eds.): H-WORKLOAD 2018, CCIS 1012, pp. 239–257, 2019.
https://doi.org/10.1007/978-3-030-14273-5_14

Thus, mental workload can influence the well being and health of a person. Furthermore, it can influence the individual performance because of forgetfulness, negligence, and a lack of concentration. The consequences are increased errors or inadequate decisions and might affect not only the own safety but also the safety of other persons. This is particularly true in safety-critical occupations such as air traffic control.

In order to understand workload changes in the air traffic sector, it is important to study the influence of different factors altering air traffic controllers' mental workload. In this article, we concentrated on two exposure parameters: air traffic volume and occurrence of an exceptional event. We were interested to find out if both of them have an effect on the experienced workload and if there was an interaction between both. In particular, we wanted to investigate if the occurrence of on exceptional event affected workload differently related to the current air traffic demands but also to individual characteristics. This understanding is relevant in order to improve working conditions by maintaining an appropriate level of workload that allows also the handling of unforeseen events.

In Sect. 2 we give a brief overview about the concept and current methods for registering mental workload. We introduce our hypotheses, the study design, and the way we proceeded for analyzing our data in Sect. 3. Finally, we outline and discuss our results in Sect. 4 as well as give prospects for future work related to our conclusions in Sect. 5.

2 Related Work

In general, mental workload was related to information processing theory [14]. High mental workload may arise from the inability to cope with increasing demands imposed on an individual's cognitive capacity [7,14,30] but also from a simultaneous interaction with emotional aspects [1], training and experience level [32]. Hence, increasing demands could originate among others from time pressure, task complexity, and individual's psycho-physiological state [11]. Methods for registering mental workload are categorized into subjective and objective methods. The subjective measurements use traditional questionnaires in order to assess subject's experienced workload. The objective methods are subdivided into performance measurement and biosignal registration. Recording and analysis of e.g. the brain activity [22,26], cardiovascular parameter [18,27] as well as ocular data [2] offered insight into subject's psycho-physiological state. The main idea underlying the assessment of workload using biosignals considered arousal and activation mechanisms of the organism reacting to the task load [21]. Measurement of individual's performance on a task was another way to determine workload. Hereby, identification of workload relied on the relationship concept between the two and implied that individual performance decreases under high mental workload [17,19,33]. However, studies also indicated that motivation, training, and experience could contribute to maintain performance at the same level by investing more effort and in this way mitigated the impact of workload [16,23]. Thus, the increased mental workload could not always be measured

directly by means of performance break-down [12,20,29]. For a detailed overview of the mental workload literature including definitions and measuring methods of workload we advise the reader on the articles of Cain [5], Vidulich and Tsang [28] as well as Stanton et al. [25].

We conclude that identification of workload by means of performance measurements was problematic whereas physiological indicators and subjective ratings using questionnaires may better reflect workload changes. However, subjective measurements were problematic because of their susceptibility to subjective distortion, social desirability restrictions regarding the appropriateness of the answer, and subject's inability to introspect. Their main advantages were the simplicity of assessment and high user acceptance.

3 Design and Methodology

3.1 Research Questions

In our simulator study, we focused on air traffic controllers as an occupation with high cognitive demands and responsibility [6]. Air traffic controllers are dealing with safety-critical tasks and have to keep engaged and try to maintain their performance even under difficult situations. When task demands increase, they have to invest more effort. As a consequence, air traffic controllers work in a high pressure environment with high mental workload. This is mainly induced by the traffic load situation itself but might also arise by unexpected events [1]. Hence, the aim of our study was to find out if it is the number of aircraft or the occurrence of an exceptional event that stresses controllers the most. Furthermore, we were interested if there is an interaction effect between both and if between-subject factors, i.e. age or job demands, could reveal additional information about controller's experienced workload in both conditions. To this end, we formulated the following five research hypotheses:

1. The number of aircraft has a significant main effect on controllers' workload.
2. The occurrence of an exceptional event has a significant main effect on controllers' workload.
3. There is a significant interaction effect between number of aircraft and occurrence of an exceptional event regarding controllers' workload.
4. The experienced workload is related to controller's age.
5. The experienced workload is related to controller's current job demands.

3.2 Traffic Scenarios

Our research was performed in the Air Traffic Management and Operations simulator (ATMOS) of the German Aerospace Center (DLR) in Braunschweig. For our research design, we concentrated on arrival management tasks and manipulated the factors: exceptional event and traffic load. The traffic load was manipulated by the number of aircraft per hour (ac/h). We considered four levels of traffic flow that determined the more or less constant number of aircraft in the

arrival sector (i.e., possible fluctuations according to controller's guiding behavior): 25 ac/h, 35 ac/h, 45 ac/h, and 55 ac/h. The second factor was by nature dichotomous: occurrence vs. absence of an exceptional event. The exceptional event was a flight that should be prioritized because of a sick passenger on board (in the following referred to as priority-flight event). The pseudo pilot was instructed to request priority for his flight but not to declare emergency by using the commands mayday or pan-pan. The rationale behind this was that in case of a mayday or pan-pan call there might be specific prescribed regulations that have to be implemented by the controller such as closing the sector, maintain a distance around the aircraft, or distribute the remaining aircraft on further controllers. These regulations would corrupt our experiment, in particular mitigate the air traffic demand factor. We decided to use the medical event communicated as priority request in order to get a workload increase in the sequence without activating additional measures which would be applicable in case of aircraft's engine failure or loss of controllability.

The combination of both factors, number of aircraft and priority-flight event, resulted in eight scenarios (Table 1). Scenario duration was 20 min for a scenario with no priority-flight event and 25 min for a scenarios with a priority-flight event. The priority-flight event occurred after the 10th min. The time parameters were chosen because of previous experiences related to simulator experiments with air traffic controllers. We gave controllers 10 min to get started and accustomed to their sector in order to control for any additional intrinsic, workload-relevant factors that could interfere with our exposure parameters. We assumed that in case of a priority-flight event the controllers would need maximally 10 min to solve it. We also knew that controllers' experience in the simulator used to be real and pervasive. By giving them 5 additional minutes in scenarios with priority-flight event, we aimed to allow them to leave the experiment with a positive impression and not with a bad feeling.

Table 1. Experimental design with two factors: number of aircraft and priority-flight event.

Simulation scenario	Traffic load (ac/h)	Priority-flight event
1	25	No
2	25	Yes
3	35	No
4	35	Yes
5	45	No
6	45	Yes
7	55	No
8	55	Yes

3.3 Procedure and Subjects

Our sample consisted of 21 subjects between the ages of 22 and 64 years (2 female, 19 male, mean age 38 ± 11). We had 13 approach controllers, 3 tower controllers, and 5 employees of the DLR, in the following referred to as novices. In real work life, subjects were working at different airports and different work positions. Thus, they had experienced different job demands. All of them had adequate expertise to handle the arrival management simulation and interact with the pseudo pilots who simulated the cockpit crews during the trials. Within two consecutive days, the subjects completed the above-mentioned eight traffic scenarios in randomized order. The first day started at noon with an introductory session where participants completed demographic questionnaires. They were briefed regarding the research goals, experimental procedure of the following two days, and workload scales used. Next, subjects completed a training session in order to get familiarized with the simulator and the questionnaires. Once they had a clear understanding of how everything worked and what was being measured, the experiment started. Four of the simulation scenarios were presented on the first day, the remaining four were conducted on the second day until noon. The Federal Institute of Occupational Safety and Health (BAuA) in Berlin was in charge of the project. All of the investigations acquired were approved by the local review board of the BAuA and the experiments were conducted in accordance with the Declaration of Helsinki. All procedures were carried out with the adequate understanding and written consent of the subjects.

3.4 Assessment of Workload

As dependent variable we assessed the experienced workload by means of the NASA-TLX, instantaneous self-assessment (ISA) questionnaire, and expert ratings. For the sake of completeness, we include the German versions used in the Appendix A.

NASA-TLX. Subjective workload was captured with a computerized version of the NASA-TLX [10]. After the training scenario, subjects were asked to rate the workload sources in 15 pairwise comparisons of NASA-TLX's six workload dimensions: mental demand, physical demand, temporal demand, performance, effort, frustration. Thereby, subjects chose the more relevant dimension of their workload. Thus, we got an individual weighting of the NASA-TLX subscales. After each simulation scenario subjects were asked to rate the scenario itself within a 100-point range regarding each of the six subscales. They indicated their rating by clicking on a 5-point step box of the scale. Finally, individual weightings S_d of the NASA-TLX dimensions d were combined with dimensions' ratings R_d according to Eq. 1 and yielded the overall workload index W_{idx} of the NASA-TLX [10].

$$W_{\text{idx}} = \frac{1}{15} \cdot \sum_{d=1}^{6} S_d \cdot R_d \tag{1}$$

ISA. During all eight scenarios controllers performed the ISA questionnaire that was developed for the assessment of air traffic controller's mental workload [4,13,15]. The ISA questionnaire consisted of a one-dimensional scale and was quick and easy to asses. It was presented in an interval of 5 min and subjects indicated their workload using a touch screen. Thereby, they had to select one of the following five values according to their feeling during the previous minutes: (1) under-utilized, (2) relaxed, (3) comfortable, (4) high, and (5) excessive. For our analysis, we only considered controller's rating after the possible occurrence of the priority-flight event, i.e., the rating of the 15th min.

Expert Ratings. At the end of each scenario, we asked the involved pseudo pilots from the simulated cockpit crews to rate the workload level of the air traffic controller during the scenario. The rating was conducted using the ISA scale. In order to have the same understanding of scale's levels as the air traffic controllers, pseudo pilots were previously briefed regarding the meaning of each level. Finally, ratings of the pseudo pilots were averaged for each scenario and participant.

3.5 Statistical Analysis

In order to answer our first three research questions regarding the effect of traffic flow, occurrence of an exceptional event, and interaction effect between both, we carried out three analysis of variance (ANOVA). The dependent variable of each was the workload index measured either with NASA-TLX, ISA, or expert ratings. For each ANOVA we utilized a repeated-measures design with two within-subject factors (two levels for the priority-flight event factor and four levels for the traffic-load factor). General differences between the levels were examined and tested with a post-hoc test (Bonferroni corrected). For testing the differences between priority-flight and no priority-flight event on each traffic-load level, we used four t-tests for each workload index and adjusted the values accordingly.

The research questions concerning group differences were examined using six mixed-factorial ANOVAs. Three of them were carried out with air traffic controller's age as between-subject factor and three with air traffic controller's current job demands. The dependent variable, within-subject factors, and levels were identical with those mentioned above. Similarly, we utilized a repeated-measures design and examined the differences with post-hoc tests (Bonferroni). In order to cluster the subjects in two groups by age, we took the median age of our sample. This yielded 11 subjects under 40 years (referred to as young) and 10 subjects over or equal 40 years (referred to as old). Work demand clustering in two groups was done by consideration of the airport traffic volume where the controller was working. Thereby, we took into account the annual report on the air transport by the DLR [3] and set a threshold in order to get two equally sized subject groups (Fig. 1). This resulted in 11 subjects working in busy airports (approach controllers or tower controllers) and 10 subjects working in smaller, less-busy airports (approach controllers, tower controllers, or novices). Finally,

we have to note that 5 subjects had to be discarded from expert rating analysis because of missing values. Hence, expert rating ANOVAs were carried out with 9 subjects in the busy-airport group and 7 subjects in the less-busy airport group and respectively, 7 young and 9 older subjects.

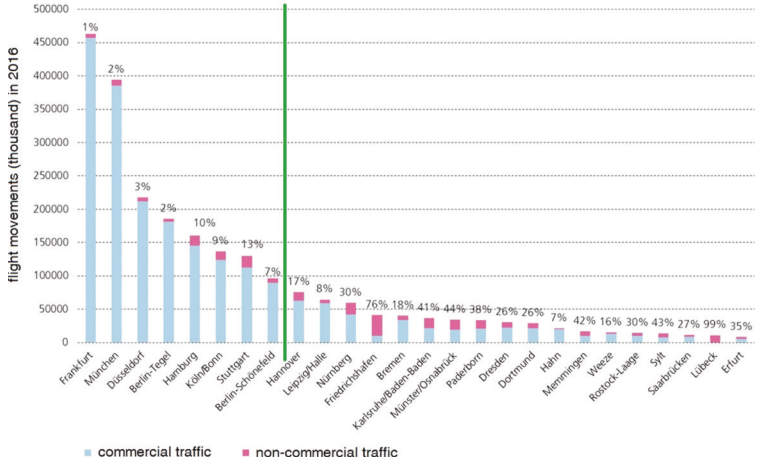


Fig. 1. Number of flight movements and share of non-commercial traffic per airport in Germany indicating our two-group split as green vertical line (figure from [3], p. 55). (Color figure online)

4 Results and Discussion

4.1 Effect of Traffic Load and Priority-Flight Event

Results of the ANOVAs for NASA-TLX, ISA, and expert ratings, each with the two within-subject factors traffic-load and priority-flight event, are summarized in Table 2.

Regarding traffic load, Bonferroni corrected post-hoc tests showed significant differences between all levels for all three measuring methods. Figure 2 shows the results. Our first hypothesis related to the effect of number of aircraft on controllers' workload proved to be true.

The impact of a priority-flight event varied across the questionnaire methods. Controllers' ISA and NASA-TLX ratings showed only week significant differences between sessions with and without priority flight. Expert ratings yielded a highly significant difference for the priority-flight event factor. In order to evaluate the effect of the priority flight for each traffic-load level, we computed t-tests and adjusted the values by means of Bonferroni correction. For expert and ISA ratings, we identified a significant difference between scenarios with and without priority-flight event for the 45 ac/h condition (experts: $t(15) = -4.28$, $p = 0.003$; ISA: $t(20) = -3.21$, $p = 0.018$). None of the other t-tests could reach significance. Our second hypothesis about the effect of an exceptional event on controllers'

Table 2. Analysis of workload scores across simulation conditions.

		F	p	η^2	Power[a]	Power[b]
Traffic load	NASA-TLX	68.224	.001	.773	.997	.414
	ISA	70.630	.001	.779	.920	.237
	Expert ratings	67.145	.001	.817	.672	.145
Priority-flight event	NASA-TLX	4.381	.049	.180	1	.998
	ISA	4.773	.041	.193	.994	.431
	Expert ratings	17.143	.001	.533	1	.684
Traffic load and priority-flight event	NASA-TLX	1.477	.230	.069	.986	.329
	ISA	1.031	.385	.049	.841	.195
	Expert ratings	1.800[c]	.183	.107	.612	.135

Note. Values of .001 are actually $p \leq 0.001$.
[a]Power indicates the a posteriori power of our study to detect medium-size effects.
[b]Power indicates the a posteriori power of our study to detect small-size effects.
[c]Indicates Mauchly's test of sphericity was significant ($p < 0.05$) and a Greenhouse-Geisser correction was made to degrees of freedom.

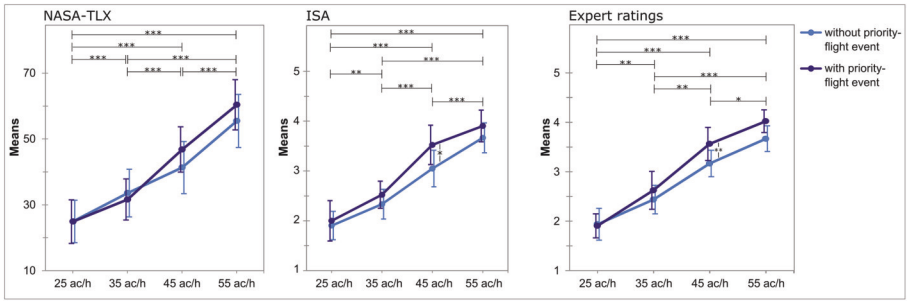


Fig. 2. Average workload over 21 participants measured using NASA-TLX (left), ISA (center), and expert ratings (right) across simulation conditions (Bonferroni corrected post-hoc tests: ∗∗∗: $p \leq 0.001$; ∗∗: $0.001 < p \leq 0.01$; ∗: $0.01 < p \leq 0.05$; error bars indicate 95% confidence interval).

workload remained unclear, in particular regarding the 25, 35, and 55 ac/h scenarios and the ratings from the NASA-TLX questionnaire. Finally, no interaction effect could be obtained between both factors with any of the questionnaires used and our third hypothesis must be refused.

4.2 Effect of Age and Job Demands

No significant effect of age could be obtained with any of the questionnaire methods (NASA-TLX: $p = 0.627$, $\eta^2 = 0.013$; ISA: $p = 0.134$, $\eta^2 = 0.114$; expert ratings: $p = 0.398$, $\eta^2 = 0.051$;). The experienced workload was not related to controller's age and thus, our fourth hypothesis has to be rejected. However, by

comparing the results descriptively (Fig. 3) we could assume that older subjects rated their workload slightly higher than younger ones during simulation scenarios with priority-flight event and higher traffic load. This held true for all three subjective measurement methods.

Regarding the between-subject factor of job demands, no significant main effect could be found using expert-rating values ($F(1, 19) = 3.188$, $p = 0.096$, $\eta^2 = 0.185$). Figure 4 (right column of top and bottom rows) shows the results scenarios with and without priority-flight event separately on two rows and indicates that it is hard to recognize a general tendency between the groups using experts' ratings.

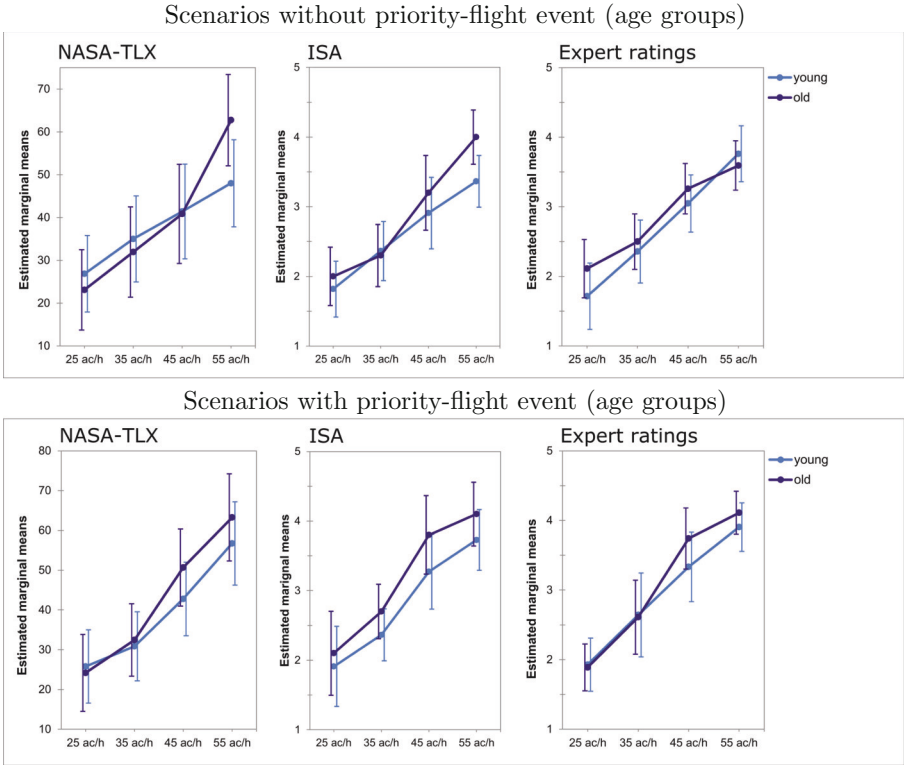


Fig. 3. Comparison of age groups. Average workload (left: NASA-TLX, center: ISA, right: expert ratings) during scenarios without (top row) and with (bottom row) priority-flight event at different traffic loads for young (in light blue) and older (in blue) subjects. Error bars indicate 95% confidence interval. (Color figure online)

Workload ratings of the subjects themselves seemed more indicative. The ISA ratings showed a significant main effect of the job demand factor ($F(1, 19) = 12.221$, $p = 0.002$, $\eta^2 = 0.391$). Although there was a significant difference in the workload means of the two groups averaged across all simulation conditions, we did not obtain a significant interaction effect of job demands with any within-subject factor. Descriptive evaluation of the results in Fig. 4 (middle column of

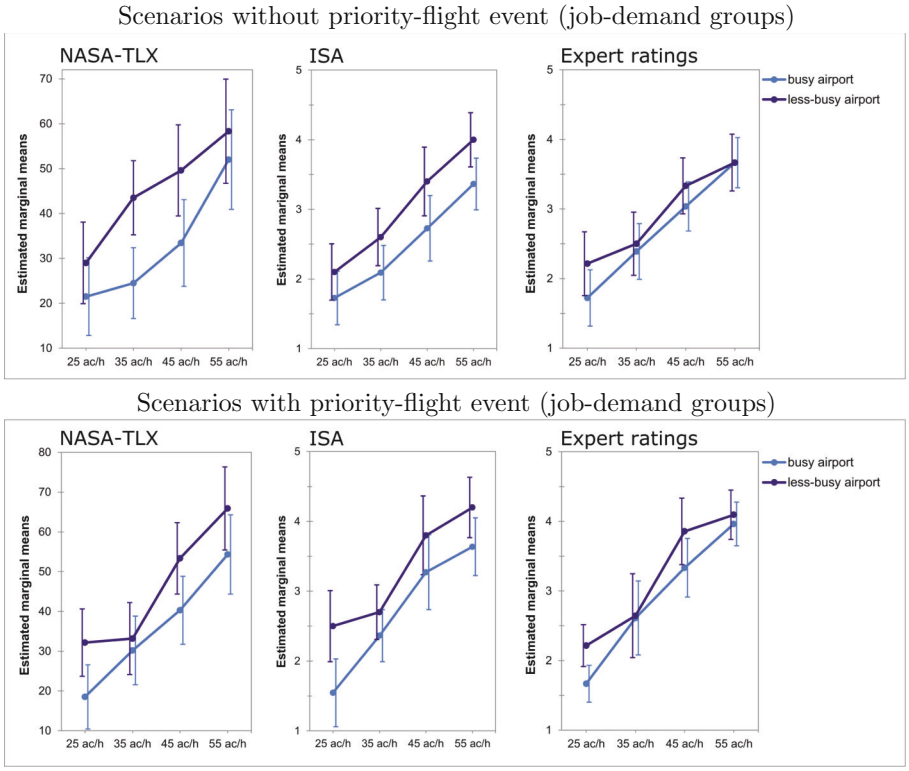


Fig. 4. Comparison of job-demand groups. Average workload (left: NASA-TLX, center: ISA, right: expert ratings) during scenarios without (top row) and with (bottom row) priority-flight event at different traffic loads for subjects working in busy (in light blue) vs. less-busy (in blue) airports. Error bars indicate 95% confidence interval. (Color figure online)

top and bottom rows) showed that the effects might be additive, meaning that the effect of job demands was similar on each traffic-load condition and the effect of traffic load was similar for each subject group. Results of conditions with and without priority-flight event were comparable.

Correspondingly, the NASA-TLX scores revealed also a significant main effect of current job demands ($F(1, 19) = 5.314$, $p = 0.033$, $\eta^2 = 0.219$). Furthermore, we were able to obtain a significant interaction between priority-flight event, traffic flow, and whether the controller was used to high job demands or not ($F(1, 57) = 3.319$, $p = 0.026$, $\eta^2 = 0.246$). The nature of this interaction is shown in Fig. 4 (left column of top and bottom rows).

In general, the subjective workload of the group working in busy airports increased gradually but to a lesser extent than the workload of the subjects working in less-busy airports. Furthermore, while workload of the subjects that work in busy airports increased almost exponentially over the traffic load of conditions without priority-flight event (i.e., between 45 ac/h to 55 ac/ah), subjective workload of subjects from less-busy airports seemed to increase somehow

logarithmically (i.e., more pronounced slope between 25 ac/h and 35 ac/h and to a lesser extent between the following traffic-load scenarios). This tendency was not prominent during the priority-flight event conditions. Hereby, subjects from less-busy airports reported the same amount of workload between 25 ac/h and 35 ac/h whereas subjects from busy airports reported a gradual increase. The experienced workload as reported by the controllers was related to controllers' current job demands and our fifth hypothesis proved to be true.

5 Conclusion and Future Work

The aim of our study was to find out if it was the number of aircraft or the occurrence of an exceptional event that stressed controllers the most and if there was an interaction effect between both. Furthermore, we were interested if the factors age and experienced job demands during real-work life could reveal additional information about controllers workload. The subjective workload was measured using NASA-TLX, ISA, and expert ratings. The simulator experiment was expected to be representative for real operations because of its similarity to controllers' working environment and controllers' communication with pseudo pilots. While the effect of the number of aircraft was evident by all three questionnaire methods, the impact of the priority-flight event remained doubtful. Controllers' ISA and NASA-TLX ratings showed only a weekly significant discrimination between sessions with and without priority flight using standard ANOVA tests. Expert ratings yielded a highly significant difference for the priority-flight event factor, in particular during the 45 ac/h scenario. An additional model based in-depth analysis of controllers' ratings using a priori assumptions on non-linear dependencies considering resource limitations might modify this conclusion.

The examination of between-subject factors could reveal additional information about controller's rating behavior. Thereby, we observed the tendency that older subjects seemed to experience more workload than younger ones, in particular during the high-traffic conditions with priority-flight event. However, the differences did not reach significance. One reason for this could be the small number of participants. In order to have two equally-sized subject groups, we took the median age of our sample as threshold. For revealing differences between age groups the threshold age should be over 40 years. Regarding the factor of job demands, assumed by the number of flight movements of the airport where subjects were working for, we obtained significant differences by means of subjects' workload ratings but not using experts' ratings. Subjects working in busy airports seemed to experience lower workload compared to the group from less-busy airports. The effect of job demands was similar on each traffic-load level and the effect of traffic load was similar for each subject group. In the main, this held true for conditions with and without priority-flight event conducted using the ISA questionnaire. Thus, we assumed that the effect of job demands may be additive. Interestingly, only the NASA-TLX scores revealed an interaction between all three factors. During scenarios without priority-flight event, we observed that workload of subjects from busy airports increased slower among low traffic-load conditions and jumped to a higher value at the highest traffic-load condition. We

suggested that experienced subjects realized their high-workload state suddenly when the traffic-load got the maximum value. In contrary, subjective workload of subjects working in less-busy airports increased most abruptly between the lowest and next-higher traffic-load condition and slower among higher traffic-load scenarios. It seemed that as subjective-workload ceiling approached the subjects of the less-busy airport group rated more cautiously. However, we have to note that this subject group included 5 (out of 10) subjects with no work experience in real airport environment. Thus, this difference between the groups could reflect the difference between novice and experienced subjects. Interestingly, subjects rated their workload more consciously during scenarios with priority-flight event.

To sum up, the number of aircraft contributed most to subjects' experienced workload while the priority-flight event became workload relevant only under high-traffic load. This observation fits well to controllers' reports. Most of them mentioned that during the scenarios with low and medium traffic volumes they had no difficulties to deal with the priority request. During the scenarios with higher traffic demands the situation changed and the priority-flight event became more demanding. Only few controllers were able to easily handle the situation. Regarding the effect of age, we conclude that more research with older controllers is necessary in order to gain more insight. Finally, the current job demands and thus, controllers habituation on higher workload states deserves more attention. The abrupt increase of perceived workload in controllers working in busy airports appears critical, in particular observed using the NASA-TLX during the high-load scenarios without priority-flight event. Objective registration of workload using bio signals may reveal if it is the workload itself that increases suddenly or if it is a lack of self-awareness that leads to these self-ratings. In this context, we want also to emphasize the importance of critical validation of metrics of mental workload as stated by [31].

Acknowledgments. We would like to thank Kerstin Ruta for her daily operational support, Emilia Cheladze for conducting the experiments, the numerous pseudo pilots for their contribution during the experiments, and Thorsten Mühlhausen for his conceptual, technical, and overall support. We would also like to thank Martin Schütte for his general project support.

More information about the project that acquired our data can be found at http://www.baua.de/DE/Aufgaben/Forschung/Forschungsprojekte/f2402.html.

Author Contributions. T.R. initiated the project and was responsible for the overall conception of the investigation. T.R., A.T., and N.F. developed the research design of the study. A.T. was responsible for the implementation of the simulation scenarios and the overall technical support. L.R. conducted the experiments and acquired the data. L.R. provided computational support for the data analysis with SPSS and graphic editing. The study was supervised by T.R. Data interpretation was performed by T.R. and B.M. The manuscript was written by T.R. Final critical editing was performed by A.T., N.F., and B.M.

Conflict of Interest Statement. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

A NASA-TLX Questionnaire

Bundesanstalt für Arbeitsschutz und Arbeitsmedizin FB 3, Gruppe 3.4

ID:_____

NASA-TLX zur Erfassung der Arbeitsbeanspruchung

Geben Sie bitte an, welche relative Bedeutung für die empfundene Gesamtbeanspruchung bei der eben durchgeführten Aufgabe die sechs folgenden Beanspruchungsdimensionen

Geistige Anforderung
Körperliche Anforderung
Zeitliche Anforderung
Ausführung der Aufgaben
Anstrengung
Frustration

für Sie hatten. Lesen Sie sich bitte zuvor bezüglich der Bedeutung der Beanspruchungsdimensionen die unten stehende Beschreibung durch und wenden Sie sich bitte bei Unklarheiten an den Versuchsleiter.

Geistige Anforderungen	Wie viel geistige Anstrengung war bei der Informationsaufnahme und bei der Informationsverarbeitung erforderlich (z.B. Denken, Entscheiden, Rechnen, Erinnern, Hinsehen, Suchen)? War die Aufgabe leicht oder anspruchsvoll, einfach oder komplex, erfordert sie hohe Genauigkeit oder ist sie fehlertolerant?
Körperliche Anforderungen	Wie viel körperliche Aktivität war erforderlich (z.B. ziehen, drücken, drehen, steuern, aktivieren …)? War die Aufgabe leicht oder schwer, einfach oder anstrengend, erholsam oder mühselig?
Zeitliche Anforderungen	Wie viel Zeitdruck empfanden Sie hinsichtlich der Häufigkeit oder dem Takt mit dem Aufgaben oder Aufgabenelemente auftraten? War die Abfolge langsam und geruhsam oder schnell und hektisch?
Ausführung der Aufgaben	Wie erfolgreich haben Sie Ihrer Meinung nach die vom Versuchsleiter (oder Ihnen selbst) gesetzten Ziele erreicht? Wie zufrieden waren Sie mit Ihrer Leistung bei der Verfolgung dieser Ziele?
Anstrengung	Wie hart mussten Sie arbeiten, um Ihren Grad an Aufgabenerfüllung zu erreichen?
Frustration	Wie unsicher, entmutigt, irritiert, gestresst und verärgert (versus sicher, bestätigt, zufrieden, entspannt und zufrieden mit sich selbst) fühlten Sie sich während der Aufgabe?

Teil 1 - Beanspruchungsstruktur

Im Folgenden werden jeweils zwei der sechs Beanspruchungsdimensionen in verschiedenen Kombinationen gegenübergestellt. Geben Sie jeweils an, welche Beanspruchungsdimension für die Gesamtbeanspruchung, die Sie empfunden haben, bedeutsamer war. Es geht also zunächst nicht darum, wie hoch Sie die Beanspruchung in der einzelnen Dimension empfanden, sondern wie wichtig die jeweilige Dimension für das Gesamtempfinden war.

Beispiel:
Wenn für Sie die geistigen Anforderungen, die die Aufgabe gestellt hat, bedeutsamer für das Beanspruchungserleben waren, als die zeitlichen Anforderungen, die Sie empfanden, kreuzen Sie bitte so an:

Geistige Anforderungen	X	☐	Zeitliche Anforderungen

Körperliche Anforderungen	☐	☐	Zeitliche Anforderungen
Anstrengung	☐	☐	Geistige Anforderungen
Frustration	☐	☐	Körperliche Anforderungen
Anstrengung	☐	☐	Frustration
Geistige Anforderungen	☐	☐	Zeitliche Anforderungen
Körperliche Anforderungen	☐	☐	Anstrengung
Zeitliche Anforderungen	☐	☐	Ausführung der Ausgaben
Frustration	☐	☐	Geistige Anforderungen
Zeitliche Anforderungen	☐	☐	Frustration
Ausführung der Aufgaben	☐	☐	Anstrengung
Anstrengung	☐	☐	Zeitliche Anforderungen
Frustration	☐	☐	Ausführung der Aufgaben
Ausführung der Aufgaben	☐	☐	Körperliche Anforderungen
Geistige Anforderungen	☐	☐	Ausführung der Aufgaben
Geistige Anforderungen	☐	☐	Körperliche Anforderungen

Kontrollieren Sie bitte, ob Sie kein Vergleichspaar vergessen haben.

Teil 2 - Beanspruchungshöhe

Geben Sie jetzt bitte an, wie hoch die Beanspruchung in den einzelnen Dimensionen bezogen auf die gerade gelöste Aufgabe ausgeprägt war. Markieren Sie dazu auf den folgenden Skalen bitte, in welchem Maße Sie sich in den sechs genannten Dimensionen von der Aufgabe beansprucht oder gefordert gesehen haben. Machen Sie dafür, wie im Beispiel dargestellt, Ihre Einschätzung mit einem Kreuz kenntlich. Bei Unklarheiten wenden Sie sich bitte an den anwesenden Versuchsleiter.

Beispiel:

Geistige Anforderungen

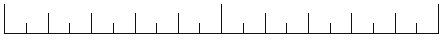
Körperliche Anforderungen

Zeitliche Anforderungen

Ausführung der Aufgaben

Anstrengung

Frustration



B ISA Questionnaire

Bundesanstalt für Arbeitsschutz und Arbeitsmedizin FB 3, Gruppe 3.4

ID:_____

ISA zur Erfassung der Arbeitsbeanspruchung

Geben Sie jetzt bitte an, wie hoch Ihre Beanspruchung in den letzten 5 Minuten war. Machen Sie dafür Ihre Einschätzung mit einem Kreuz kenntlich. Bei Unklarheiten wenden Sie sich bitte an den anwesenden Versuchsleiter.

				
Der Lotse hat sehr wenig bis gar nichts zu tun.	Der Lotse hat mehr als erforderlich, um die Aufgaben zu erfüllen. Die Zeit vergeht langsam.	Der Lotse hat ausreichend Arbeit. Alle Aufgaben sind unter Kontrolle.	Der Lotse ist „am Limit". Bestimmte nicht-zwingend notwendige Aufgaben werden verschoben. Die Zeit vergeht schnell	Der Lotse ist überlastet. Einige Aufgaben können nicht erledigt werden. Der Lotse spürt, dass er nicht die Kontrolle hat

ID:..... | Datum: ..... | VL: ..... Seite **1** von **2**

References

1. Averty, P., Collet, C., Dittmar, A., Athènes, S., Vernet-Maury, E.: Mental workload in air traffic control: an index constructed from field tests. Aviat. Space Environ. Med. **75**(4), 333–341 (2004)
2. Beatty, J.: Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychol. Bull. **91**(2), 276–292 (1982)
3. Berster, P., et al.: Luftverkehrsbericht 2016 - Daten und Kommentierungen des deutschen und weltweiten Luftverkehrs (2017). http://www.dlr.de/fw/Portaldata/42/Resources/dokumente/pdf/__Luftverkehrsbericht_2016_final_141217.pdf
4. Brennan, S.: An experimental report on rating scale descriptior sets for the instantaneous self assessment (ISA) recorder. Technical report. DRA Technical Memorandum (CAD5) 92017, DRA Maritime Command and Control Division (1992)
5. Cain, B.: A review of the mental workload literature. Technical report, Defence Research and Development Canada, Human System Integration Section, Toronto (2007). http://ftp.rta.nato.int/public/pubfulltext/rto/tr/rto-tr-hfm-121-part-ii/tr-hfm-121-part-ii-04.pdf. Accessed 11 Dec 2013
6. Edwards, T., Martin, L., Bienert, N., Mercer, J.: The relationship between workload and performance in air traffic control: exploring the influence of levels of automation and variation in task demand. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 120–139. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_8
7. Eggemeier, F., Wilson, G.F., Kramer, A.F., Damos, D.L.: Workload assessment in multi-task environments. In: Multiple-task Performance, pp. 207–216. Taylor & Francis (1991)

8. Hancock, P.A.: Whither workload? Mapping a path for its future development. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 3–17. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_1
9. Hancock, P.A., Desmond, P.A.: Stress, Workload, and Fatigue. Lawrence Erlbaum Associates Publishers, Hillsdale (2001)
10. Hart, S.G., Staveland, L.E.: Development of the NASA TLX: results of empirical and theoretical research. In: Hancock, P., Meshkati, N. (eds.) Human Mental Workload, pp. 139–183, North Holland, Amsterdam (1988)
11. Hendy, K.C., Hamilton, K.M., Landry, L.N.: Measuring subjective workload: when is one scale better than many? Hum. Factors **35**, 579–601 (1993)
12. Hilburn, B., Jorna, P.G.A.M.: Workload and air traffic control. In: Hancock, A., Desmond, P. (eds.) Stress, Workload, and Fatigue, pp. 384–394. Erlbaum, Mahwah (2001)
13. Jordan, C.: Experimental study of the effect of an instantaneous self assessment workload recorder on task performance. Technical report. DRA Technical Memorandum (CAD5) 92011, DRA Maritime Command Control Division (1992)
14. Kahneman, D.: Attention and Effort. Prentice-Hall, Englewood Cliffs (1973)
15. Kirwan, B., et al.: Human factors in the ATM system design life cycle. In: FAA/Eurocontrol ATM R&D Seminar (1997)
16. Matthews, G.: Levels of transaction: a cognitive sciences framework for operator stress, pp. 5–33. Lawrence Erlbaum Associates Publishers Mahwah (2001)
17. Meister, D.: Behavioral Foundations of System Development. Wiley, New York (1976)
18. Mulder, L.J.M., van Roon, A.M., Althaus, M., Laumann, K., Dicke, M.: Determining dynamic cardiovascular state changes using a baro-reflex simulation model. In: Human factors in transportation, communication, health and the workplace, pp. 297–315. Shaker Publishing, Maastricht (2002)
19. O'Donnell, C.R.D., Eggemeier, F.T.: Workload assessment methodology. In: Boff, K., Kaufmann, L., Thomas, J. (eds.) Handbook of Perception and Human Performance, pp. 42–49. Wiley (1989). http://apps.usd.edu/coglab/schieber/docs/odonnell.pdf, Accessed 19 June 2015
20. Parasuraman, R., Hancock, P.: Adaptive control of workload. In: Hancock, A., Desmond, P. (eds.) Stress, Workload, and Fatigue, pp. 305–320. Erlbaum, Mahwah (2001)
21. Pribram, K.H., McGuinness, D.: Arousal, activation, and effort in the control of attention. Psychol. Rev. **82**, 116–149 (1975). http://www.karlpribram.com/wp-content/uploads/pdf/theory/T-068.pdf. Accessed on 15 May 2015
22. Radüntz, T.: Dual frequency head maps: a new method for indexing mental workload continuously during execution of cognitive tasks. Front. Physiol. **8**, 1019 (2017). https://www.frontiersin.org/article/10.3389/fphys.2017.01019
23. Saxby, D.J., Matthews, G., Warm, J.S., Hitchcock, E.M., Neubauer, C.: Active and passive fatigue in simulated driving: discriminating styles of workload regulation and their safety impacts. J. Exp. Psychol.: Appl. **19**(4), 287–300 (2013)
24. Sonnentag, S., Kruel, U.: Psychological detachment from work during off-job time: the role of job stressors, job involvement, and recovery-related self-efficacy. Eur. J. Work. Organ. Psychol. **15**(2), 197–217 (2006). https://doi.org/10.1080/13594320500513939
25. Stanton, N., Salmon, P., Walker, G., Baber, C., Jenkins, D.: Human Factors Methods: A Practical Guide for Engineering and Design, December 2005

26. Ullsperger, P., Metz, A.M., Gille, H.G.: The P300 component of the event-related brain potential and mental effort. Ergonomics **31**(8), 1127–1137 (1988). https://doi.org/10.1080/00140138808966752. pMID: 3191898
27. Veltman, J.A., Gaillard, A.W.K.: Physiological indices of workload in a simulated flight task. Biol. Psychol. **42**(3), 323–342 (1996). http://www.sciencedirect.com/science/article/pii/0301051195051651
28. Vidulich, A.M., Tsang, P.: Mental workload and situation awareness, pp. 243–273, 4th edn. Wiley, Hoboken (2012)
29. de Waard, D.: The measurement of drivers' mental workload. Ph.D. thesis, University of Groningen, Traffic Research Centre, Haren, Netherlands (1996)
30. Wickens, C.D.: Multiple resources and performance prediction. Theor. Issues Ergon. Sci. **3**(2), 159–177 (2002). http://www.tandfonline.com/doi/abs/10.1080/14639220210123806
31. Wickens, C.D.: Mental workload: assessment, prediction and consequences. In: Longo, L., Leva, M.C. (eds.) H-WORKLOAD 2017. CCIS, vol. 726, pp. 18–29. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61061-0_2
32. Xie, B., Salvendy, G.: Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments. Work Stress **14**(1), 74–99 (2000)
33. Yerkes, R.M., Dodson, J.D.: The relation of strength of stimulus to rapidity of habit-formation. J. Comp. Neurol. Psychol. **18**, 459–482 (1908). http://psychclassics.yorku.ca/Yerkes/Law. Accessed 03 Nov 2011

frontiers in Physiology

ORIGINAL RESEARCH
published: 21 April 2020
doi: 10.3389/fphys.2020.00300

Indexing Mental Workload During Simulated Air Traffic Control Tasks by Means of Dual Frequency Head Maps

Thea Radüntz[1], Norbert Fürstenau[2], Thorsten Mühlhausen[2] and Beate Meffert[3]*

[1] Mental Health and Cognitive Capacity, Work and Health, Federal Institute for Occupational Safety and Health, Berlin, Germany, [2] Institute of Flight Guidance, German Aerospace Center, Braunschweig, Germany, [3] Signal Processing and Pattern Recognition, Department of Computer Science, Humboldt-Universität zu Berlin, Berlin, Germany

In our digitized society, advanced information and communication technology and highly interactive work environments impose high demands on cognitive capacity. Optimal workload conditions are important for assuring employee's health and safety of other persons. This is particularly relevant in safety-critical occupations, such as air traffic control. For measuring mental workload using the EEG, we have developed the method of Dual Frequency Head Maps (DFHM). The method was tested and validated already under laboratory conditions. However, validation of the method regarding reliability and reproducibility of results under realistic settings and real world scenarios was still required. In our study, we examined 21 air traffic controllers during arrival management tasks. Mental workload variations were achieved by simulation scenarios with different number of aircraft and the occurrence of a priority-flight request as an exceptional event. The workload was assessed using the EEG-based DFHM-workload index and instantaneous self-assessment questionnaire. The DFHM-workload index gave stable results with highly significant correlations between scenarios with similar traffic-load conditions (r between 0.671 and 0.809, $p \leq 0.001$). For subjects reporting that they experienced workload variation between the different scenarios, the DFHM-workload index yielded significant differences between traffic-load levels and priority-flight request conditions. For subjects who did not report to experience workload variations between the scenarios, the DFHM-workload index did not yield any significant differences for any of the factors. We currently conclude that the DFHM-workload index reveals potential for applications outside the laboratory and yields stable results without retraining of the classifiers neither regarding new subjects nor new tasks.

Keywords: mental workload, psychophysiology, air traffic controllers, electroencephalography, biomedical signal processing, pattern recognition, state monitoring

OPEN ACCESS

Edited by:
Ahsan H. Khandoker, Khalifa University, United Arab Emirates

Reviewed by:
Gianluca Di Flumeri, Sapienza University of Rome, Italy
Pietro Aricò, Sapienza University of Rome, Italy

****Correspondence:***
Thea Radüntz
raduentz.thea@baua.bund.de

Specialty section:
This article was submitted to Computational Physiology and Medicine, a section of the journal Frontiers in Physiology

Received: *29 May 2019*
Accepted: *17 March 2020*
Published: *21 April 2020*

Citation:
Radüntz T, Fürstenau N, Mühlhausen T and Meffert B (2020) Indexing Mental Workload During Simulated Air Traffic Control Tasks by Means of Dual Frequency Head Maps. Front. Physiol. 11:300. doi: 10.3389/fphys.2020.00300

1. INTRODUCTION

In our digitized society, advanced information and communication technology and highly interactive work environments impose high demands on cognitive capacity and on the ability to cope with increased task load (Kompier and Kristensen, 2001; Niosh, 2002; Landsbergis et al., 2003; Lohmann-Haislah, 2012). According to several authors mental workload can be conceived as the amount of cognitive demands required in order to solve a task related to the cognitive resources available (Kahneman, 1973; Eggemeier et al., 1991; Xie and Salvendy, 2000; Wickens, 2002).

Optimal workload conditions are important for the health of the single individual and in order to assure the safety of other persons. Latter is particularly relevant in safety-critical occupations with high cognitive demands and responsibility, such as air traffic control. A valid and reliable method for measuring mental workload would offer a way for achieving such conditions in human-machine systems by capturing the instantaneous workload continuously over time (Byrne and Parasuraman, 1996; Scerbo et al., 2001; Arico et al., 2017). It is important that the registration method does not interact with the task or alter subject's mental state by imposing additional demands as it is the case during subjective assessment of workload by means of questionnaires. Furthermore, the workload should not only be detectable in retrospect or after the occurrence of errors as it is the case when performance measures are used for workload detection. Thus, questionnaires and performance evaluation are only of limited relevance for real-time analysis of workload conditions in the range of seconds.

Over the past 50 years, various physiological parameters (e.g., heart rate and derived parameters, electrodermal activity, body temperature, etc.) have been evaluated for their validity regarding continuous mental workload registration. Since the discovery of the electroencephalogram (EEG) by Berger (1929), relations between bioelectric brain activity and cognitive states have been studied. Improvements of the amplifier technology and computerized evaluation of biosignals made systematic investigations possible. In last century's 90s, the state-of-the art regarding EEG's evaluation and validity was summarized in reviews that served as a starting point for the use of the EEG in applied research, e.g., in human-factors. In a review article, Borghini et al. (2014) provided a detailed overview of the measurement of neurophysiological signals for the determination of mental workload and confirmed essentially the known relations. The authors further concluded that no convincing algorithms were available for a reliable online workload detection.

The spectral power of oscillations in different frequency bands were used as parameters for describing the spontaneous brain activity. For the alpha-frequency (8–12 Hz) and theta-frequency (4–8 Hz) bands, spectral-power comparisons in all relevant investigations described systematic relations to cognitive and memory performance (Sterman and Mann, 1995; Pfurtscheller, 1997; Gevins et al., 1998; Klimesch, 1999; Gevins and Smith, 2000). These EEG bands were also linked to different levels of workload by means of analysis of variance (e.g., Mecklinger et al., 1992; McEvoy et al., 2001; Lei and Roetting, 2011; Brouwer et al., 2012; Capilla et al., 2012; Aricò et al., 2018) and demonstrated a decrease of the alpha-frequency band power and an increase of the theta-frequency band power with increasing mental workload.

In recent years, however, classifiers were increasingly used for the separation of workload levels. The feature vectors—derived from the EEG—revealed varying complexity and extent, and frequency bands were taken differently into account. The used EEG parameters were, for example, the amplitude of the EEG signal, spectral power of different frequency bands, and different EEG channels (Wilson and Russell, 2003b; Lin et al., 2006; Kohlmorgen et al., 2007; Baldwin and Penaranda, 2012; Penaranda and Baldwin, 2012; Ke et al., 2014). The focus was on frontal, parietal and occipital EEG channels according to previous findings. Independent component analysis (ICA) was used to determine specific reactions of spatio-temporal different sources (Gardony et al., 2017) and allowed the successful detection and elimination of artifacts (Mognon et al., 2011; Radüntz et al., 2017; Puma et al., 2018).

Initially, studies that dealt with the determination of workload were conducted in the laboratory using different task batteries (Gevins et al., 1998; Gevins and Smith, 2000; McEvoy et al., 2001; Berka et al., 2007; Grimes et al., 2008; Baldwin and Penaranda, 2012; Brouwer et al., 2012, 2014; Christensen and Estepp, 2013; Weiland et al., 2013; Gerjets et al., 2014; Hogervorst et al., 2014; Ke et al., 2014; Hou et al., 2016; Gardony et al., 2017; Rosen and Reiner, 2017; Puma et al., 2018). Meanwhile, investigations of cognitive workload with more realistic tasks became more popular (Kohlmorgen et al., 2007; Lei and Roetting, 2011; Aricò et al., 2018; Dehais et al., 2018). Air traffic controllers (ATCOs) pose a special challenge due to the complex task-load situations with changing activities and strategies for air traffic management (ATM). The requirements can change very fast, a clear and direct objective graduation of task-load proves to be difficult, and the transitions are often unpredictable and fast. Experiments with ATM simulations and a task-load grading proved to be advantageous although the majority of simulated ATM examinations were limited to two task-load levels (easy and difficult). Relevant studies on workload determination methods for simulated or real air traffic control were conducted by Brookings et al. (1996), Wilson and Russell (2003b, 2007), Shou et al. (2012), Abbass et al. (2014b,c), Borghini et al. (2014, 2017), Aricò et al. (2015, 2016), Aricò et al. (2018), Di Flumeri et al. (2015), Dasari et al. (2017), and Dehais et al. (2018).

Wilson and Russell (2003a) investigated the classification of the mental state of seven air traffic controllers in simulated air traffic monitoring. In seven different task-load conditions a 19-channel EEG, heart rate, blink rate, and respiratory rate were recorded. The spectral power of five frequency bands was calculated for each EEG channel from 1-s windows and used per subject as input for the artificial neuronal networks (ANN) and stepwise linear discriminant analysis (SWLDA). Discrimination only between two conditions yielded the best result with an accuracy of 97.5% (ANN) and 91% (SWLDA). Thereby only 22 relevant features were included in the evaluation. The authors drew attention to the following open questions of day-to-day variability of psychophysiological measures and long training duration for ANN. They stated that a one-size-fits-all solution would be beneficial.

Abbass et al. (2014c) dealt with questions about visual and auditory information processing in relation to mental workload of air traffic controllers. In addition, the authors examined the question of whether a narrow-band frequency resolution of the EEG was better suited for the assessment of workload. They found that there were no quantitative advantages over the usual frequency bands. Further, they suggested to focus on the separation of high and low workload and neglect the middle range (Abbass et al., 2014a).

The question of reliability of EEG-based workload determination in ATM tasks was examined in Arico et al. (2015). According to the authors the reason for the decreasing classification accuracy over days, as reported by Christensen and Estepp (2013), could be overfitting, i.e., a too high specificity of the training data. It was hypothesized that a simple classifier based on fewer spectral properties guaranteed a high selectivity over days. Twelve ATCO interns completed the simulated ATM task on 2 consecutive days and after 9 days. The EEG was registered by 13-channels and 2 s windows were used to compute relevant EEG spectral features. For each subject, cross-validation of the classifier between the days was calculated using 5, 50, and 100% of relevant EEG features. The results showed that the use of only 5% of the relevant features contributed to an over-day stable workload measurement.

Basically, changes in the alpha-frequency and theta-frequency band powers related to mental workload have been confirmed many times and proved to be meaningful in accordance with the findings of the last 50 years. The majority of workload studies dealt with the analysis of the EEG during cognitive tasks related to working memory and executive control. While some authors investigated whether a brain-state monitoring was possible on the basis of universal and general activation signs in the EEG (Bashivan et al., 2014, 2015; Ke et al., 2014), others tested the possibilities and limitations of over task requirements (cross-task training) and inter-individually (cross-subject training) transferable classifiers. Discrimination accuracy of the classifiers between high and low workload was often not sufficient in cross-task training and remained below the significance threshold. Cross-subject training of the classifier was also less favorable than intra-subject classification. In the driving simulator study by Kohlmorgen et al. (2007), the authors concluded that a highly adaptive approach was needed to account for the neurophysiological variations. According to the authors, a universally applicable "workload detector" with fixed parameters did not seem to be realistic at the moment. The selection of appropriate data for classifier's training needs more elucidation. This is especially important as frequent allegations were made concerning the time interval between training and test of the classifier that proved to be particular relevant for the classification accuracy (Penaranda and Baldwin, 2012). In order to avoid overfitting and increase the stability of the classifier performance over time a smaller number of features could be beneficial (Arico et al., 2015).

It has to be stated that different cognitive strategies in task solving, both intra- and inter-individually, can influence the classification results. In this context, Puma et al. (2018) suggested to cluster the subjects according to their performance, age (McEvoy et al., 2001), and individual experiences. These should be considered if workload registration methods are to be validated.

Based on the possibility that machine learning algorithms provide the ability of workload registration in the range of seconds, the question arises whether they provide reliable and reproducible results over time, in particular without the need for re-training of the classifier regarding subjects and tasks. For their practical application at the workplace, it is also important that their applicability is examined not only in the laboratory but also under more realistic conditions. This becomes particularly important when considering the technological advancements regarding mobile EEG technology that have simplified signal registration outside of shielded rooms (Mihajlovic et al., 2015; Aricò et al., 2018; Radüntz, 2018; Baek et al., 2019; Radüntz and Meffert, 2019).

In our prior work we developed a mental-workload classifier that does not need retraining, neither for new subjects nor for new tasks (Radüntz, 2017). In a laboratory study conducted with 54 subjects and during execution of well-established cognitive tasks, we developed the so-called Dual Frequency Head Maps (DFHM). These head maps consist of personalized spectral features and their spatial occurrence (i.e., frontal theta-band and parietal alpha-band powers). Support vector machines are used for classification in three classes: low, moderate, or high workload. Under laboratory conditions, we successfully proved the DFHM method as universally applicable with fixed parameters for mental-workload indexing. For proofing the reliability and reproducibility of our DFHM method's results under realistic conditions, we conducted a study in cooperation with the German Aerospace Center and focused on air traffic controllers. The following four research hypotheses were formulated for the DFHM-validation study:

1. The DFHM method yields stable results under similar task-load conditions independently of the time of measurement.
2. The DFHM method is able to assess workload differences that arise from different traffic-volumes conditions.
3. The DFHM method is able to assess workload differences that arise from an exceptional-event condition.
4. The objectively measured workload assessed by the DFHM method is related to controller's subjectively experienced workload.

2. MATERIALS AND METHODS

2.1. Research Design

Our study took place at the Air Traffic Management and Operations Simulator (ATMOS) of the German Aerospace Center (DLR) in Braunschweig. Thereby, air traffic controllers focused on simulated arrival management procedures presented on the monitor and interacted along the experimental task with pseudo pilots who simulated the cockpit crews. The implemented simulation scenarios differed regarding two factors that were responsible for mental workload variations of air traffic controllers as suggested by Averty et al. (2004). The first one was the traffic load. In our case, we had four levels corresponding to four different numbers of aircraft per hour (ac/h). The second factor was an exceptional event that could occur or not. This event was a pilot's request for a flight prioritization because of a sick passenger on board. The priority-flight request could occur around the 11th min of the 20–25 min lasting scenario. Both factors led to the eight scenarios presented in **Table 1**.

TABLE 1 | Independent variables and simulation scenarios.

		Number of aircraft per hour			
		Low (25 ac/h)	**Medium (35 ac/h)**	**High (45 ac/h)**	**Very high (55 ac/h)**
Exceptional event	No	Scenario 1	Scenario 3	Scenario 5	Scenario 7
	Yes	Scenario 2	Scenario 4	Scenario 6	Scenario 8

TABLE 2 | Experimental procedure.

Duration (min)	**Procedure**	
	Day 1: ca. 12.30–17.30	**Day 2: ca. 9.30–12.30**
120	Briefing, training	
65	Two simulation scenarios	Two simulation scenarios
15	Break	Break
65	Two simulation scenarios	Two simulation scenarios

2.2. Procedure and Subjects

We asked subjects to participate in a 2-days experiment where they had to complete the above-mentioned eight traffic scenarios in randomized order. The experimental procedure is outlined in **Table 2**. The investigation consisted of an introductory session and the main experiment. During the introductory session participants completed demographic questionnaires, were briefed regarding the research goals and experimental procedure of the following 2 days, and had a training session at the simulator in order to get familiarized with the environment. During the main experiment the subjects completed four of the simulation scenarios while the remaining four were conducted on the second day.

In our study, we examined 21 subjects in the age between 22 and 64 years (2 females, 19 males, mean age 38 $\pm$ 11). Subjects were from different airports, had different work experience, revealed different work positions (i.e., 13 approach controllers, three tower controllers, and five employees of the DLR), and had experienced different work demands. However, all of them had adequate expertise to handle the arrival management simulation.

All of the investigations acquired were approved by the local review board of our institution and complied with the tenets of the Declaration of Helsinki. All procedures were carried out with the adequate understanding and written consent of the subjects.

2.3. Subjective Ratings

In order to register the subjectively experienced workload, we used the instantaneous self-assessment (ISA) questionnaire. This was developed for the assessment of air traffic controller's mental workload (Brennan, 1992; Jordan, 1992; Kirwan et al., 1997) and consisted of a one-dimensional scale. Thus, it was quickly and easily conducted in an interval of 5 min during all eight scenarios. According to their feeling during the previous 5 min, subjects indicated their workload using a touch screen. Thereby, they selected one of the following five values: (1) under-utilized, (2) relaxed, (3) comfortable, (4) high, and (5) excessive.

Analysis of the ISA questionnaire results was particularly relevant for our fourth hypothesis related to controller's subjectively experienced workload. Based on these we developed a so-called workload-sensitivity index that considered the individual range of experienced workload during different task-load conditions.

Subject's normalized workload-sensitivity index s_a was based on a linear model for the dependence of subjectively experienced workload as assessed by the ISA questionnaire and traffic load. In Fürstenau et al. (2020), we showed that the linear model was able to predict the ISA value with a high confidence for means across the subjects and provided reasonable linear correlation coefficients for the individuals. Independence from the arbitrary ISA values was achieved via normalization by scales' means, i.e., $(\text{traffic load}_{max} + \text{traffic load}_{min})/2$ for the traffic volume and $(ISA_{max} + ISA_{min})/2$ for the subjective workload, resulting in anticorrelated (normalized) sensitivity and intercept $s_b = 1 - s_a$. ISA-scale means were conducted individually for each subject based on the ISA-extreme values from their regression lines.

Our workload-sensitivity index ranged between 0.32 and 1.23, and was used for subject clustering. The aim of this clustering was an improved investigation of the cognitive phenomena only of those subjects that actually experienced different workload levels. Subjects with an index below the median of 0.8 were clustered as not sensitive, while subjects with an index equal or above the median as workload sensitive. Generally speaking, workload-sensitive subjects experienced more workload variation during the different simulation scenarios whereas the not-sensitive subjects rated the subjectively experienced workload with less variation.

2.4. EEG and DFHM-Workload Index

Biosignal processing and all calculations were done with MATLAB.

For EEG registration we used g.tec's g.LADYbird/g.Nautilus system with 25 active electrodes placed at positions according to the 10–20-system (**Figure 1**). Registration was carried out with a sample rate of 500 Hz and with reference to electrode Cz. For signal recording we used g.tec's Matlab interface.

After recording, the EEG was filtered with a bandpass filter (order 100) between 0.5 and 40 Hz for enhancing the separation accuracy of the following analysis for artifact rejection (Fernandez, 2009; Omatu et al., 2010; Pignat et al., 2013; Winkler et al., 2015). Independent component analysis [ICA, Infomax algorithm (Makeig et al., 1996)] for artifact rejection was applied to the signal. Components to reject were manually selected (i.e., on average 16 out of 25 per subject). In order to

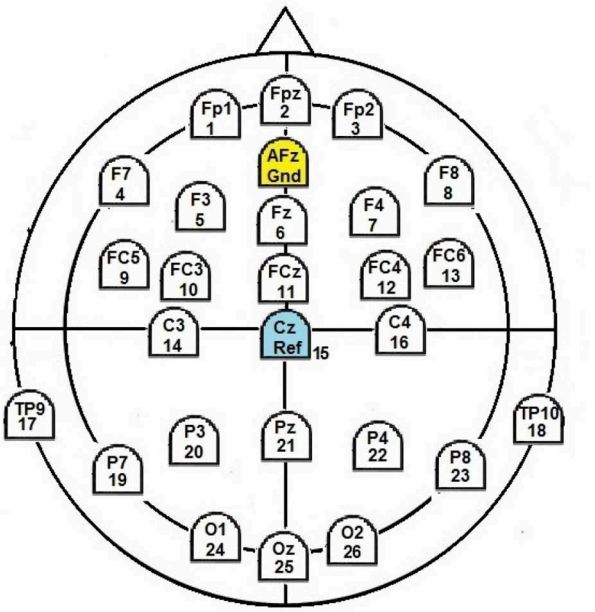


FIGURE 1 | EEG layout used.

TABLE 3 | Mean and standard deviation (in parenthesis) of the α and θ frequency band powers exemplary for two electrodes averaged over the subjects for each simulation scenario.

Number of aircraft per hour	25 ac/h	35 ac/h	45 ac/h	55 ac/h
Without an exceptional event				
θ frequency band power (Fz electrode)	16.5 (4.0)	17.5 (4.3)	17.4 (3.7)	18.2 (3.8)
α frequency band power (Pz electrode)	26.1 (6.1)	25.2 (5.4)	25.4 (5.7)	25.1 (5.2)
With an exceptional event				
θ frequency band power (Fz electrode)	16.7 (3.7)	17.4 (4.1)	17.3 (3.7)	17.8 (3.9)
α frequency band power (Pz electrode)	26.0 (6.0)	25.5 (5.6)	25.0 (5.2)	24.9 (4.4)

increase topographical localization, we applied a simple Hjorth-style surface laplacian filter using eight neighbors (Hjorth, 1975). This spatial high-pass filter was aimed to attenuate large-scale scalp signals and amplify localized signals.

The artifact-free EEG was transformed to average reference and cut into segments of 1 s length, overlapping by 0.5 s. By means of Fast Fourier Transformation (FFT) we computed the workload relevant frequency bands (theta: 4–8 Hz, alpha: 8–12 Hz) over the segments. **Table 3** shows the general tendencies of both frequency bands exemplary for two electrodes. Involvement of all electrodes, the combination of both frequency bands, and the personalization of the band-power values aim at enhancing workload classification and constitute the DFHM that were generated as outlined in Radüntz (2017). In brief, we applied a theta-bandpass filter to the signals of the frontal electrodes and an alpha-bandpass filter to the signals of the parietal electrodes and calculated for each participant, each electrode, and each segment the z-scores of theta and alpha band power. The individual mean and standard deviation for z-score calculation were obtained from subject's segments of the first minute of each scenario. This compilation of the z-scores of the theta band power from the frontal electrodes and alpha band power from the parietal electrodes constituted the DFHM for each EEG segment. Next, each DFHM from the simulation scenarios' segments was classified using the already trained SVM classifier from the laboratory study. Retraining of the DFHM classifier was not necessary neither for the new subjects nor for the new tasks. The general characteristic of these maps and thus, the classifier is universally applicable because of the z-score calculation. For more information about the DFHM and classifier development, we refer the interested reader to our method article (Radüntz, 2017).

We obtained every 0.5 s a value of 1 (low workload), 2 (moderate workload), or 3 (high workload). We applied a moving-average time window of 30 s as suggested by Abbass et al. (2014a) and adjusted the result in order to gain a DFHM-workload index between 0 and 100 (Equation 1; with t: workload index at time t, DFHM (i): classification value of DFHM from segment i).

$$\mathrm{WKLindex(t)} = \Big(\sum_{i=t-59}^{t} DFHM(i) - 60\Big)/120 * 100 \qquad (1)$$

In particular, for each moving-average time window of 30 s we firstly calculated the sum of the 60 values resulting from the DFHM every 0.5 s. In order to have a baseline of 0, we subtracted the minimum-possible sum of 60 for the case where all DFHM of the window indicated a low workload of 1. Thus, the maximum-possible sum for the case where all DFHM of the window indicated a high workload was 120. Dividing by the latter and multiplying by 100 provided the percentage amount of high-workload segments in a time-window of 30 s. This constituted the DFHM-workload index between 0 and 100 computed every 0.5 s.

2.5. Statistical Analysis

For evaluating our first hypothesis and proof the reliability of the DFHM index, we calculated the DFHM-index average over the first 5 min of each simulation and correlated the means of scenarios with same traffic load.

For investigating the ability of the DFHM method to assess mental workload arising from the traffic volume (hypothesis 2) and the occurrence of an exceptional event (hypothesis 3), we looked at the time slots immediately after the time of a possible priority-flight request. This was triggered in the data using g.tec's g.TRIGbox. For scenarios with a priority-flight request we considered a DFHM-index segment of 2.5 min starting from the request time point. For scenarios without a priority-flight request we used the same time slots. We carried out an analysis of variance (ANOVA) with the slots' mean DFHM index as dependent variable. We utilized a repeated-measures design with two within-subject factors (two levels for the priority-flight request factor and four levels for the traffic-volume factor). General differences between the levels were examined and tested with a *post-hoc* test (Bonferroni corrected).

TABLE 4 | Correlation analysis of DFHM-index means over the time slot 0–5 min during scenarios with equal traffic-load volume ($N = 21$, $^{***}p \leq 0.001$).

	Traffic load			
	25 ac/h	**35 ac/h**	**45 ac/h**	**55 ac/h**
Pearson's correlation coefficient	0.671***	0.809***	0.798***	0.746***

Finally, we addressed the issue of DFHM-index workload registration in relation to subjects' subjectively experienced workload (hypothesis 4). We clustered our subjects in two groups using the median of our workload-sensitivity index that was calculated from the ISA ratings. This yielded nine subjects that subjectively did not experience workload variations between the scenarios and 12 workload-sensitive subjects. We carried out a mixed ANOVA with cluster affiliation as between-subject factor followed by a two-factorial ANOVA for each cluster separately for determining the simple main effects of our factors. The dependent variable, within-subject factors, and levels were identical with those mentioned above. Similarly, we utilized a repeated-measures design and examined the differences with *post-hoc* tests (Bonferroni).

Statistical calculations were conducted using SPSS and the significance threshold was set at 5%.

3. RESULTS

3.1. DFHM Index Under Similar Conditions

Our first hypothesis was concerned with the ability of the DFHM method to yield stable results under similar task-load conditions. Scenarios with and without priority-flight request were identical regarding their traffic volumes until the 10th min where the request could occur. Thus, we decided to use only the first 5 min of each simulation for assuring similar task load conditions between both values to be correlated. By taking a larger slot, the scenarios would increasingly differ the more time passed away as consequence of the interactive communication of the ATC with the pseudo pilots.

Correlation analyses between the mean DFHM index of the first 5 min of simulation scenarios with same traffic load showed significant positive correlations. These were particularly high for the traffic-load conditions of 35 and 45 ac/h and less pronounced for the lowest traffic load of 25 ac/h. Person's correlation coefficients are presented in **Table 4**.

3.2. DFHM Index Related to Traffic Load and Priority-Flight Request

In order to evaluate the ability of the DFHM method to assess workload differences arising from different traffic-volume and exceptional-event conditions, we considered the results of the ANOVA. They were calculated with the two within-subject factors traffic-load and priority-flight request. The results are summarized in **Table 5**.

Related to our second hypothesis the traffic load had a significant main effect on the workload as assessed by the DFHM index. Bonferroni corrected *post-hoc* tests showed significant differences between all levels except between the 35 and 45 ac/h conditions. The DFHM-workload index increased with increased traffic. **Figure 2** shows the results. The impact of the priority-flight request as related to our third hypothesis did not became significant. No interaction effect could be obtained between traffic load and priority-flight request.

TABLE 5 | Analysis of DFHM index across simulation conditions over all subjects and subjects' clusters, respectively.

		F	*p*	η^2
Traffic load	All	22.953[a]	0.001	0.534
	Workload-sensitive subjects	36.815	0.001	0.769
	Not-sensitive subjects	2.762	0.064	0.257
Priority-flight request	All	1.349	0.259	0.063
	Workload-sensitive subjects	15.636	0.002	0.587
	Not-sensitive subjects	1.311	0.285	0.141
Traffic load and priority-flight request	All	0.214	0.886	0.011
	Workload-sensitive subjects	0.936	0.434	0.078
	Not-sensitive subjects	0.440	0.726	0.052

Values of 0.001 are actually $p \leq 0.001$.
[a] Indicates Mauchly's test of sphericity was significant ($p < 0.05$) and a Greenhouse-Geisser correction was made to degrees of freedom.

For assuring that air traffic controllers indeed prioritized the aircraft, we evaluated the route distances of the same aircraft with and without priority-flight request. In both cases the route distance taken was the length of trajectory between the initial contact time point and landing. A shorter route distance for the requesting aircraft indicated that air traffic controllers complied with the priority-request condition (**Figure 3**). Wilcoxon signed-ranks tests (with Bonferroni correction) indicated that the route distance was significantly shorter during scenarios with priority-flight request compared to scenarios with same traffic volume but without priority-flight request (**Table 6**).

3.3. DFHM Index Related to Subjectively Experienced Workload Variations

For our last hypothesis, results from the mixed ANOVA showed statistically significant interaction effects between cluster affiliation and traffic load [$F_{(3,\ 57)} = 7.215$, $p < 0.001$, $\eta^2 = 0.275$] as well as between cluster affiliation and priority-flight request [$F_{(1,\ 19)} = 9.517$, $p = 0.006$, $\eta^2 = 0.334$]. No significant interaction effect could be obtained between all three factors cluster affiliation, traffic load, and priority-flight request [$F_{(3,\ 57)} = 1.195$, $p = 0.319$, $\eta^2 = 0.059$].

In the following, we analyzed the DFHM index for the workload-sensitive cluster and the not-sensitive cluster separately. For the workload-sensitive cluster the ANOVA yielded a significant main effect for the traffic load and priority-flight request. Bonferroni corrected *post-hoc* tests showed significant differences between all traffic-load levels except between the highest traffic load volumes with 45 and 55 ac/h. The DFHM-workload index increased with increased traffic load and was higher during scenarios with priority-flight request. No interaction effect could be obtained between both factors.

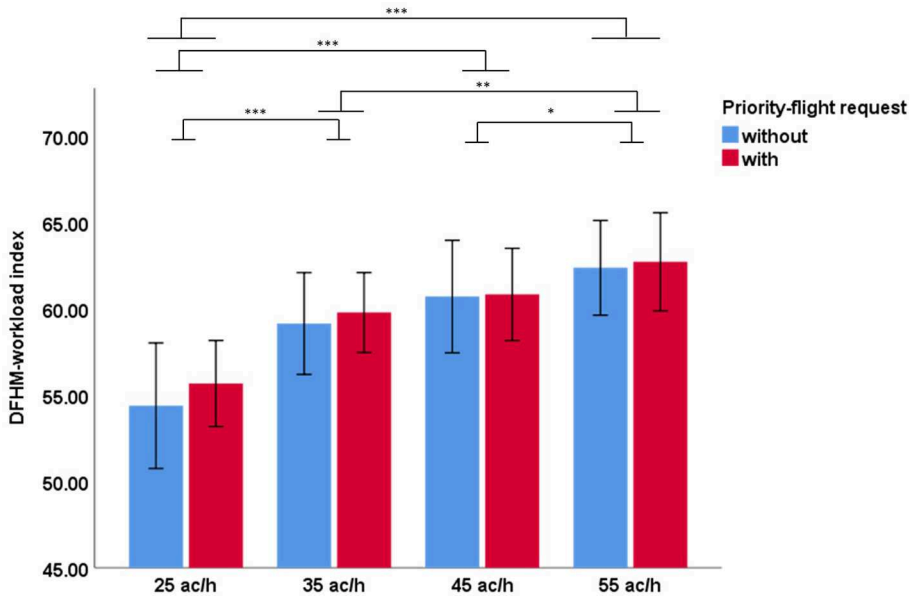


FIGURE 2 | Mean DFHM index over 21 participants measured during the 2.5 min slots after a possible priority-flight request across simulation conditions with (red) and without (blue) priority-flight request at different traffic loads (Bonferroni corrected *post-hoc* tests: ***$p \leq 0.001$; **$0.001 < p \leq 0.01$; *$0.01 < p \leq 0.05$; error bars indicate 95% confidence interval).

For the not-sensitive cluster no significant differences could be obtained for none of the factors. The results are summarized in **Table 5** and shown in **Figure 4**.

3.4. Performance Related to Subjectively Experienced Workload Variations

In addition to the DFHM index we evaluated the performance of the air traffic controllers for the workload-sensitive and not-sensitive clusters. As measure of performance we employed the route distances and loss of separation. Evaluation of route distance between aircraft with priority-flight request and without was conducted separately for each cluster. The results are presented in **Table 6** and **Figure 5** and revealed similar tendencies for both clusters, i.e., the route distance of the requesting aircraft was significantly shorter during 35 and 45 ac/h traffic load. During the 55 ac/h condition this held true only for the not-sensitive cluster. No significant difference could be found for none of the clusters during the 25 ac/h condition.

Evaluation of loss of separation between aircraft was conducted according to the minimum separation standards specified by the authorities and based on the standards of the International Civil Aviation Organization (2011). The separation minima were breached when lateral distance between two aircraft was smaller than the required vake vortex separation, i.e., 3 NM (nautical miles) for two medium type aircraft and 5 NM for a medium aircraft following a heavy aircraft, and simultaneously vertical distance between these aircraft was smaller than 1,000 ft. In general, the number of loss of separation was low (i.e., around zero) and thus not appropriate for statistical evaluation. For the sake of completeness, **Figure 6** illustrates the results for each cluster separately.

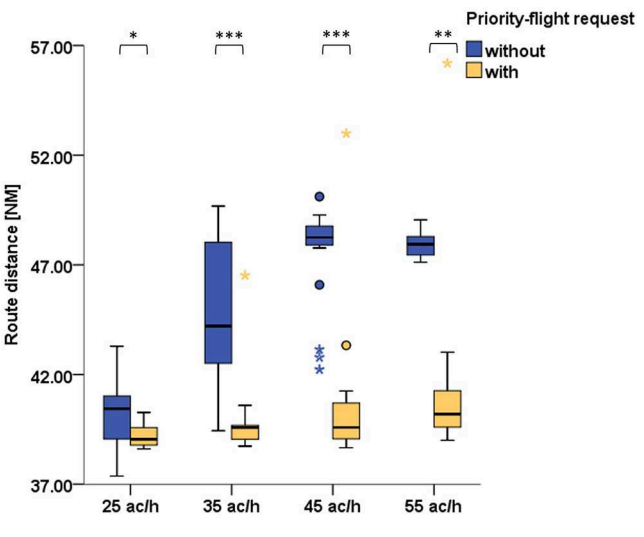


FIGURE 3 | Comparison of prioritized aircraft's route distance during scenarios with priority-flight request (orange) and during scenarios with same traffic volume but without prioritization (blue) for all 21 subjects (Wilcoxon signed-ranks tests with Bonferroni correction: ***$p \leq 0.001$; **$0.001 < p \leq 0.01$; *$0.01 < p \leq 0.05$).

4. DISCUSSION

In our study, we aimed in validating our method for mental workload registration by means of DFHM. The method was already proofed in a laboratory setting but further evaluation was needed. Our current validation study was conducted under realistic conditions, with real tasks, and new subjects, i.e., in an air traffic control simulator, with arrival-management tasks, and

TABLE 6 | Wilcoxon signed-ranks tests (with Bonferroni correction) for comparison of prioritized aircraft's route distance during scenarios with priority-flight request and aircraft's route distance during scenarios with same traffic volume but without priority-flight request.

	Median (range) route distance [NM]		*Z*	*p*	*r*
	Without priority-flight request	**With priority-flight request**			
All subjects (*N* = 21)					
25 ac/h	40.44 (5.92)	39.05 (1.66)	−2.52	0.047	−0.55
35 ac/h	44.21 (10.24)	39.58 (7.79)	−4.02	0.001	−0.88
45 ac/h	48.25 (7.88)	39.59 (14.33)	−3.84	0.001	−0.84
55 ac/h	47.94 (1.93)	40.20 (17.19)	−3.46	0.002	−0.76
Workload-sensitive subjects (*N* = 12)					
25 ac/h	40.37 (4.83)	38.95 (1.10)	−1.49	0.544	−0.43
35 ac/h	44.18 (9.93)	39.36 (1.14)	−3.06	0.009	−0.88
45 ac/h	48.03 (7.04)	39.61 (14.33)	−2.67	0.031	−0.77
55 ac/h	47.61 (1.93)	40.42 (17.19)	−2.35	0.074	−0.68
Not-sensitive subjects (*N* = 9)					
25 ac/h	40.45 (4.68)	39.52 (1.64)	−1.96	0.203	−0.65
35 ac/h	44.64 (7.17)	39.59 (7.74)	−2.67	0.031	−0.89
45 ac/h	48.74 (2.07)	39.58 (1.85)	−2.67	0.031	−0.89
55 ac/h	48.13 (1.16)	39.91 (3.97)	−2.67	0.031	−0.89

Values of 0.001 are actually $p \leq 0.001$.

air traffic controllers. Our sample set consisted of 21 subjects that completed eight simulation scenarios in randomized order. The simulation scenarios differed regarding their traffic load that consisted of four levels and a priority-flight request that could occur around the 11th min of simulation or not. We registered the EEG during the simulations and computed the DFHM-workload index for each subject and scenario. We did not retrain the classifiers neither for the new tasks nor for the new subjects. The gained results were promising.

The DFHM index gave stable results with highly significant correlations between scenarios with similar traffic-load conditions as stated by hypothesis 1. We observed that these correlations were particularly pronounced during the medium and high traffic volumes and less strong for the low-traffic volume. During the latter, requirements were very low and allowed air traffic controllers to have task-unrelated thoughts in order to cope with boredom (Cummings et al., 2015). Boredom proneness, coping strategy as well as the kind of task-unrelated thoughts could have mitigated the correlation between the two 25 ac/h scenarios. One could argue that there might be also other factors that might influence results stability across scenarios, e.g., effects of learning and fatigue in the course of time, the interaction with the pseudo pilots, or the initial excitement during the presentation of the first scenario. However, our sample was very specialized. Air traffic controllers are highly trained and it seemed unlikely that they gained knowledge in the course of the experiment. The initial training phase prior to our experiment was aimed to familiarize the subjects with the environmental conditions and eliminate issues related to these. For minimizing fatigue effects, we followed the regulations of working-time organization for air traffic controllers that prescribe a break after 120 min of work. Each scenario had a maximal duration of 25 min, a break took place after two scenarios (i.e., after 50 min), and the daily session consisted of four simulation scenarios. Effective daily-work time was 100 min the most. Hence, fatigue effects should be minimal. Presentation order of the scenarios was randomized and should compensate the initial excitement across subjects. Finally, air traffic controllers should be used to the interaction with different pilots from their daily work experience. Hence, we concluded that workload differences should result from the experimental conditions and the DFHM-workload index should be comparable during the first minutes of simulations with equal traffic load. Nevertheless, we have to draw attention on the increased requirement on our DFHM-workload index because of our 2-days experiment with randomized presentation order of the scenarios. Keeping this in mind, results from the correlation analysis appear encouraging.

While the first hypothesis was concerned with test-retest reliability, the second and third hypotheses addressed the issue of validity of the DFHM method as workload indexing technique. The DFHM index was able to assess significant differences between the different levels of air traffic volume as stated by hypothesis 2. Problematic were the neighboring levels with 35 and 45 ac/h that could not be significantly discriminated by the DFHM-workload index when considered over all subjects. The same held true regarding the priority-flight request although evaluation of the route distance of the requesting aircraft indicated that air traffic controllers complied with the task. At this stage hypothesis 3 had to be rejected when considered over all subjects.

More insight regarding intra-individual differences linked to the DFHM-workload index was gained from subject clustering by means of the subjectively experienced workload differences

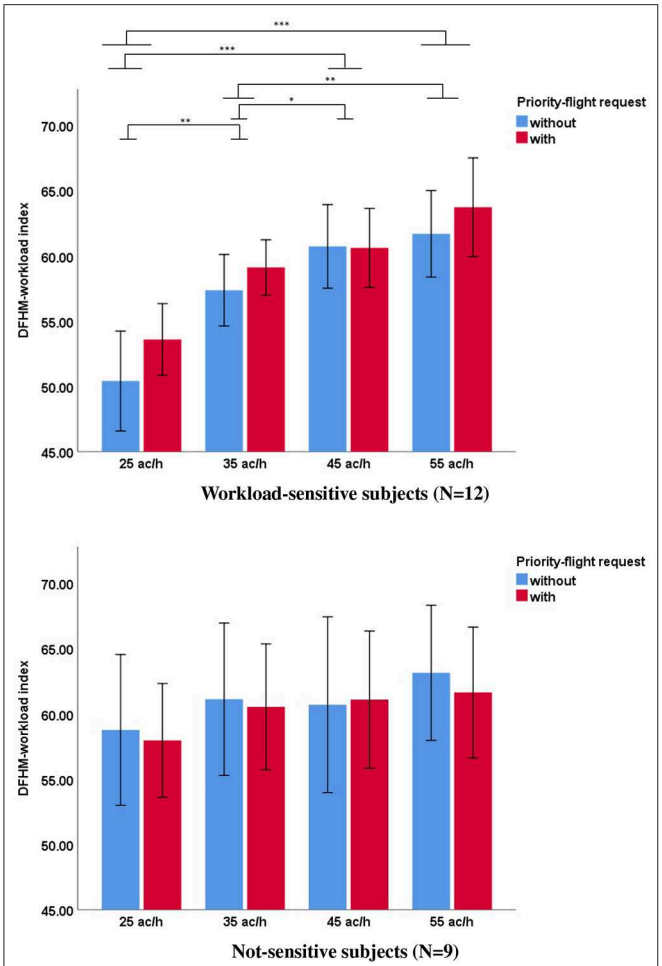


FIGURE 4 | Mean DFHM index during scenarios with (red) and without (blue) priority-flight request at different traffic loads for workload-sensitive (top row) and not-sensitive (bottom row) subjects (Bonferroni corrected *post-hoc* tests: ***$p \leq 0.001$; **$0.001 < p \leq 0.01$; *$0.01 < p \leq 0.05$; error bars indicate 95% confidence interval).

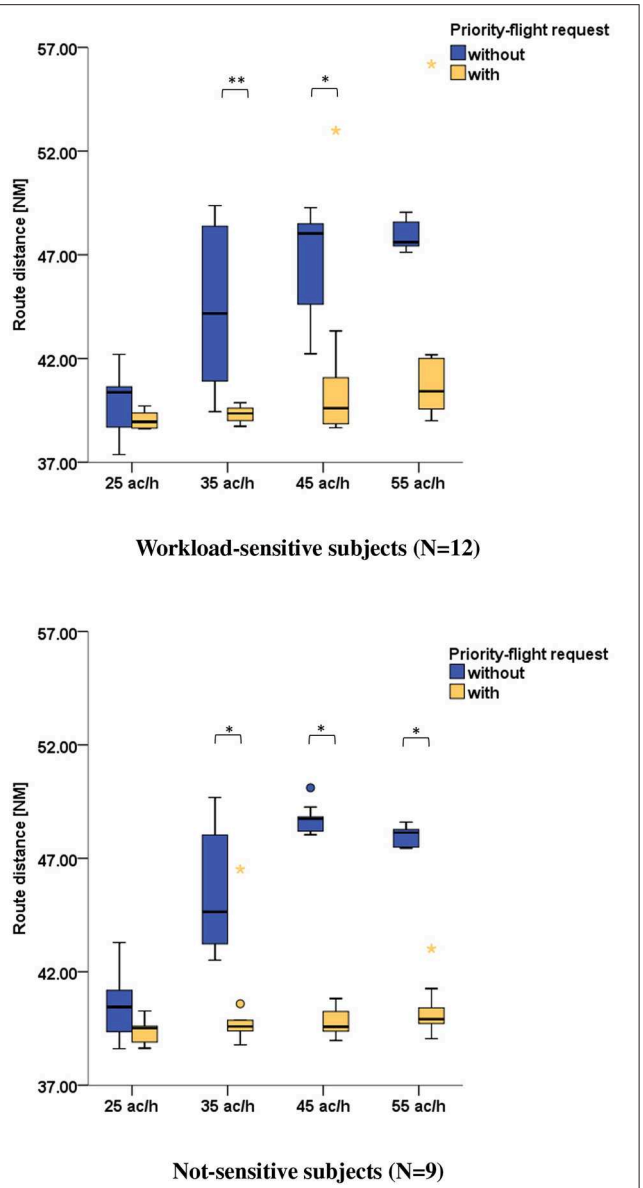


FIGURE 5 | Comparison of prioritized aircraft's route distance during scenarios with priority-flight request (orange) and during scenarios with same traffic volume but without prioritization (blue) for workload-sensitive (top row) and not-sensitive (bottom row) subjects (Wilcoxon signed-ranks tests with Bonferroni correction: **$0.001 < p \leq 0.01$; *$0.01 < p \leq 0.05$).

during the scenarios. Thus, our fourth hypothesis dealt not only with issues of validity but also of consistency between subjective and objective measuring methods. We were able to obtain highly significant interaction effects between subjective workload-cluster affiliation and traffic load as well as priority-flight request. For subjects reporting that they experienced workload variation between the different scenarios, the DFHM-workload index yielded significant differences between traffic-load levels and priority-flight request conditions. Interestingly, for these subjects the DFHM index was able to differentiate between the neighboring levels with 35 and 45 ac/h but not between the 45 and 55 ac/h conditions. Descriptive evaluation of **Figure 4** indicates that for the workload-sensitive subjects there was a ceiling effect regarding traffic volume. This occurred for traffic-volumes >45 ac/h and seemed reasonable when taken into account that a traffic volume of 55 ac/h was a condition that is highly improbable in reality for single-runway operations. Latter was constructed for the simulation in order to create an extreme situation that would definitely challenge the operators and increase their workload. Nevertheless, air traffic controllers are trained to adjust their work strategies in order to assure safety. This strategy change could be a reason for the ceiling effect during the very high traffic-load condition. However, the occurrence of a priority-flight request during the very high traffic-load condition led to a further increase of the DFHM-workload index. Unfortunately, our small sample size and the even smaller amount of subjects

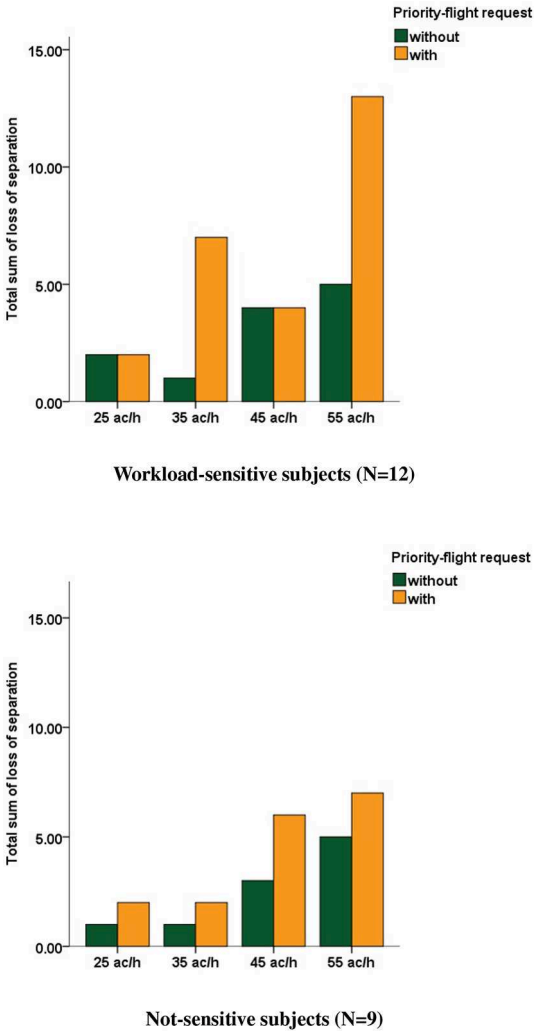


FIGURE 6 | Total sum of loss of separation computed over all workload-sensitive (top row) and not-sensitive (bottom row) subjects during scenarios with (orange) and without (green) priority-flight request at different traffic loads.

in the clusters did not allow for elaborated statistics regarding interaction effects.

In contrast to the significant differences obtained for the workload-sensitive cluster, the DFHM-workload index behaved differently for the not-sensitive cluster and did not yield any significant differences for any of the factors. In our opinion, this fit well to our fourth hypothesis and indicated that the objectively measured workload assessed by the DFHM method corresponded to controller's subjectively experienced workload. To sum up, hypothesis 2 and 3 proofed true only for subjects that experienced workload differences also subjectively during the scenarios. The workload insensitivity of subjects might appear odd when considering the high variability of our experimental design. An explanation might be traced back to the different cognitive strategies in task solving, both intra- and inter-individually, that might influence the experienced workload. Each controller had a different way to handle the traffic. This was possibly related to the different individual experience level from daily-work life as linked to the size of the airport he was working, the different ages, but also personality traits. Unfortunately, we were not able to identify personal characteristics for each cluster that might be responsible for the different perceptions of workload. More research is needed in order to understand which individual factors contribute to these interpersonal differences.

Analyses of performance data emphasized these findings. Results revealed a tendency to more loss of separation and lower prioritization during the extreme traffic load condition for workload-sensitive subjects that was less pronounced for the not-sensitive subjects. These might be an additional indicator that subjects from the workload-sensitive cluster experienced more workload compared to the others as evident by the DFHM-workload index. As a side note, readers might wonder that route-distance difference was low between the 25 ac/h scenarios with and without priority-flight request. This was reasonable because of the low-traffic volume that allowed air traffic controllers to instruct pilots to fly direct routes to the final approach even without a priority-flight request by the pilot. Conversely, a weaker significance level for the route-distance difference between both 55 ac/h scenarios could be linked to a smaller ability to prioritize the aircraft due to increased demands resulting from the high-traffic load.

A limitation of our study was the realization of the exceptional event as recurring priority-flight request. The surprising effect of the unexpected event might have diminished after the first occurrence of the request. Thereafter, air traffic controllers might have adjusted their strategy and behavior in order to be prepared to appropriately react to a recurring event. Studies that aim to understand the effect of an unexpected event on workload, should pay more attention on this issue. Finally, a larger sample size would be beneficial.

5. CONCLUSIONS

With the development and availability of low-cost and easy-to-use EEG sensors, amplifiers, and signal-processing algorithms over the last 20 years (Lopez-Gordo et al., 2014; Radüntz, 2018; Flumeri et al., 2019; Radüntz and Meffert, 2019), certain frequency bands of the EEG have proven to be particularly informative and were therefore being used more and more frequently for mental-workload detection. The numerous studies published after the year 2000 were fairly different, depending on the specific question, purpose, and expertise of the authors (Lin et al., 2006; Berka et al., 2007; Kohlmorgen et al., 2007; Borghini et al., 2014; Ke et al., 2014; Bashivan et al., 2015; Aricò et al., 2016). Initially, the spectral power in the alpha and theta frequency bands were identified as particular relevant, analyzed, and tested variance-analytically related to mental workload. In the last few years classifiers that relied on large property vectors of EEG activity were increasingly developed. Thereby, the derived parameters let barely identify the concrete psycho-physiological

meaning of the EEG activity. We aimed to avoid this issue by making use of well-established parameters that should be valid for different subjects and tasks.

In our article, we particularly addressed questions of functionality outside the laboratory, stability of results, and the generalization properties of the DFHM-workload index, inter-individually and cross-task. In conclusion, it can be stated that a reliable determination of mental workload in a realistic setting and with real-world scenarios was possible. Continuous determination under real conditions, however, requires further systematic investigations. Although the temporal resolution of the EEG permits a workload determination in the range of seconds, the states to be detected originate from long-running procedures and therefore require further research about an informative time frame for averaging classifier's output. Future promising applications of the DFHM-workload index include research about effects of human-computer interaction, human factors, ergonomic designs of the cognitive state as an objective method for development and testing new interfaces, determination of the effectiveness of training and simulation programs, or even the characterization of group dynamics when collecting synchronous EEG data from multiple subjects. The recently increasing attempts of a real-time application of EEG parameters to determine vigilance, emotion, workload, and stress are accompanied by the effort of catchy visualization of the results. With an easy accessibility of such systems, however, there is also an increasing risk of uncritical assessment and interpretation of the measured values by laymen.

DATA AVAILABILITY STATEMENT

The conducted data used to support the findings of this study are restricted by the ethics committee of the Federal Institute for Occupational Safety and Health in order to protect subjects privacy according to data-protection regulations. Data can be made available from the corresponding author upon request and after approval of the legal department for researchers who meet the criteria for access to confidential data.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Federal Institute for Occupational Safety and Health. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TR initiated the project and was responsible for the overall conception of the investigation. TR, TM, and NF developed the research design of the study. TM was responsible for the implementation of the simulation scenarios and the overall technical support. TR was responsible for the signal processing, data analysis, and method of evaluation. The study was supervised by TR. Data interpretation was performed by TR and BM. The manuscript was written by TR. Final critical editing was performed by TM, NF, and BM.

FUNDING

This study was funded by the Federal Institute for Occupational Safety and Health (project number: F 2402). More information about the project where our data were acquired can be found under the following link: https://www.baua.de/DE/Aufgaben/Forschung/Forschungsprojekte/f2402.html.

ACKNOWLEDGMENTS

We would like to thank Kerstin Ruta for her daily operational support, our student assistants Lea Rabe, Emilia Cheladze, and Friederice Schröder for conducting the experiments, the numerous pseudo pilots for their contribution during the experiments, Marion Freyer for computational support with SPSS, our student assistants Yuexin Cao and Ilona Pritschke for graphic editing, and André Tews for his conceptual, technical, and overall support. We would also like to thank Peter Ullsperger and Martin Schütte for their general project support.

REFERENCES

Abbass, H. A., Tang, J., Amin, R., Ellejmi, M., and Kirby, S. (2014a). Augmented cognition using real-time EEG-based adaptive strategies for air traffic control. *Proc. Hum. Fact. Ergon. Soc. Annu. Meeting* 58, 230–234. doi: 10.1177/1541931214581048

Abbass, H. A., Tang, J., Amin, R., Ellejmi, M., and Kirby, S. (2014b). The computational air traffic control brain: Computational red teaming and big data for real-time seamless brain-traffic integration. *J. Air Traffic Control* 52, 10–17.

Abbass, H. A., Tang, J., Ellejmi, M., and Kirby, S. (2014c). Visual and auditory reaction time for air traffic controllers using quantitative electroencephalograph (QEEG) data. *Brain Inform.* 1, 39–45. doi: 10.1007/s40708-014-0005-8

Aricó, P., Borghini, G., Di Flumeri, G., Colosimo, A., Bonelli, S., Golfetti, A., et al. (2016). Adaptive automation triggered by EEG-based mental workload index: a passive brain-computer interface application in realistic air traffic control environment. *Front. Hum. Neurosci.* 10:539. doi: 10.3389/fnhum.2016.00539

Arico, P., Borghini, G., Flumeri, G. D., Bonelli, S., Golfetti, A., Graziani, I., et al. (2017). Human factors and neurophysiological metrics in air traffic control: a critical review. *IEEE Rev. Biomed. Eng.* 10, 250–263. doi: 10.1109/RBME.2017.2694142

Arico, P., Borghini, G., Flumeri, G. D., Colosimo, A., Graziani, I., Imbert, J.-P., et al. (2015). "Reliability over time of EEG-based mental workload evaluation during air traffic management (ATM) tasks," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Milan: IEEE). doi: 10.1109/EMBC.2015.7320063

Aricó, P., Borghini, G., Flumeri, G. D., Sciaraffa, N., and Babiloni, F. (2018). Passive BCI beyond the lab: current trends and future directions. *Physiol. Meas.* 39:08TR02. doi: 10.1088/1361-6579/aad57e

Aricó, P., Borghini, G., Graziani, I., Imbert, J.-P., Granger, G., Benhacene, R., et al. (2015). Air-traffic-controllers (ATCO): neurophysiological analysis of training and workload. *Ital. J. Aerospace Med.* 2015:35.

Averty, P., Collet, C., Dittmar, A., Athénes, S., and Vernet-Maury, E. (2004). Mental workload in air traffic control: an index constructed from field tests. *Aviat. Space Environ. Med.* 75, 333–341.

Baek, H. J., Chang, M. H., Heo, J., and Park, K. S. (2019). Enhancing the usability of brain-computer interface systems. *Comput. Intell. Neurosci.* 2019, 1–12. doi: 10.1155/2019/5427154

Baldwin, C. L., and Penaranda, B. N. (2012). Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification. *Neuroimage* 59, 48–56. doi: 10.1016/j.neuroimage.2011.07.047

Bashivan, P., Bidelman, G. M., and Yeasin, M. (2014). Spectrotemporal dynamics of the EEG during working memory encoding and maintenance predicts individual behavioral capacity. *Eur. J. Neurosci.* 40, 3774–3784. doi: 10.1111/ejn.12749

Bashivan, P., Yeasin, M., and Bidelman, G. M. (2015). "Single trial prediction of normal and excessive cognitive load through EEG feature fusion," in *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* (Philadelphia, PA: IEEE). doi: 10.1109/SPMB.2015.7405422

Berger, H. (1929). Über das Elektrenkephalogramm des Menschen. *Archiv. Psychiatr. Nervenkr.* 87, 527–570. doi: 10.1007/BF01797193

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., et al. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviat. Space Environ. Med.* 78, B231–B244.

Borghini, G., Aricó, P., Di Flumeri, G., Cartocci, G., Colosimo, A., Bonelli, S., et al. (2017). EEG-based cognitive control behaviour assessment: an ecological study with professional air traffic controllers. *Sci. Rep.* 7:547. doi: 10.1038/s41598-017-00633-7

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., and Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci. Biobehav. Rev.* 44, 58–75. doi: 10.1016/j.neubiorev.2012.10.003

Brennan, S. (1992). *An Experimental Report on Rating Scale Descriptior Sets for the Instantaneous Self Assessment (ISA) Recorder*. Technical Report DRA Technical Memorandum (CAD5) 92017, DRA Maritime Command and Control Division.

Brookings, J. B., Wilson, G. F., and Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biol. Psychol.* 42, 361–377. doi: 10.1016/0301-0511(95)05167-8

Brouwer, A.-M., Hogervorst, M. A., Holewijn, M., and van Erp, J. B. F. (2014). Evidence for effects of task difficulty but not learning on neurophysiological variables associated with effort. *Int. J. Psychophysiol.* 93, 242–252. doi: 10.1016/j.ijpsycho.2014.05.004

Brouwer, A. M., Hogervorst, M. A., van Erp, J. B. F., Heffelaar, T., Zimmerman, P. H., and Oostenveld, R. (2012). Estimating workload using EEG spectral power and ERPs in the n-back task. *J. Neural Eng.* 9:045008. doi: 10.1088/1741-2560/9/4/045008

Byrne, E. A., and Parasuraman, R. (1996). Psychophysiology and adaptive automation. *Biol. Psychol.* 42, 249–268. doi: 10.1016/0301-0511(95)05161-9

Capilla, A., Schoffelen, J.-M., Paterson, G., Thut, G., and Gross, J. (2012). Dissociated α-band modulations in the dorsal and ventral visual pathways in visuospatial attention and perception. *Cereb. Cortex* 24, 550–561. doi: 10.1093/cercor/bhs343

Christensen, J. C., and Estepp, J. (2013). Coadaptive aiding and automation enhance operator performance. *Hum. Fact.* 55, 965–975. doi: 10.1177/0018720813476883

Cummings, M. L., Gao, F., and Thornburg, K. M. (2015). Boredom in the workplace. *Hum. Fact. J. Hum. Fact. Ergon. Soc.* 58, 279–300. doi: 10.1177/0018720815609503

Dasari, D., Shou, G., and Ding, L. (2017). ICA-derived EEG correlates to mental fatigue, effort, and workload in a realistically simulated air traffic control task. *Front. Neurosci.* 11:297. doi: 10.3389/fnins.2017.00297

Dehais, F., Dupres, A., Flumeri, G. D., Verdiere, K., Borghini, G., Babiloni, F., et al. (2018). "Monitoring pilot's cognitive fatigue with engagement features in simulated and actual flight conditions using an hybrid fNIRS-EEG passive BCI," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (Miyazaki: IEEE). doi: 10.1109/SMC.2018.00102

Di Flumeri, G., Borghini, G., Aricó, P., Colosimo, A., Pozzi, S., Bonelli, S., et al. (2015). "On the use of cognitive neurometric indexes in aeronautic and air traffic management environments," in *Symbiotic Interaction*, eds B. Blankertz, G. Jacucci, L. Gamberini, A. Spagnolli, and J. Freeman (Cham: Springer International Publishing), 45–56. doi: 10.1007/978-3-319-24917-9_5

Eggemeier, F., Wilson, G. F., Kramer, A. F., and Damos, D. L. (1991). *Multiple-Task Performance, Chapter Workload Assessment in Multi-Task Environments*. London: Taylor & Francis, 207–216.

Fernandez, H. Z. (2009). *Evaluation and Comparsion of the Independent Components of Simultaneously Measured MEG and EEG Data*. Berlin: Univ.-Verlag der TU.

Flumeri, G. D., Aricó, P., Borghini, G., Sciaraffa, N., Florio, A. D., and Babiloni, F. (2019). The dry revolution: evaluation of three different EEG dry electrode types in terms of signal spectral features, mental states classification and usability. *Sensors* 19:1365. doi: 10.3390/s19061365

Fürstenau, N., Radüntz, T., and Mühlhausen, T. (2020). Model-based development of a mental workload-sensitivity index for subject clustering. *Theor. Issues Ergon. Sci.* 1–25. doi: 10.1080/1463922X.2020.1711990

Gardony, A. L., Eddy, M. D., Brunyé, T. T., and Taylor, H. A. (2017). Cognitive strategies in the mental rotation task revealed by EEG spectral power. *Brain Cogn.* 118, 1–18. doi: 10.1016/j.bandc.2017.07.003

Gerjets, P., Walter, C., Rosenstiel, W., Bogdan, M., and Zander, T. O. (2014). Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Front. Neurosci.* 8:385. doi: 10.3389/fnins.2014.00385

Gevins, A., and Smith, M. E. (2000). Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style. *Cereb. Cortex* 10, 829–839. doi: 10.1093/cercor/10.9.829

Gevins, A., Smith, M. E., Leong, H., McEvoy, L., Whitfield, S., Du, R., et al. (1998). Monitoring working memory load during computer-based tasks with EEG pattern recognition methods. *Hum. Fact. J. Hum. Fact. Ergon. Soc.* 40, 79–91. doi: 10.1518/001872098779480578

Grimes, D., Tan, D. S., Hudson, S. E., Shenoy, P., and Rao, R. P. N. (2008). "Feasibility and pragmatics of classifying working memory load with an electroencephalograph," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '08* (New York, NY: ACM), 835–844.

Hjorth, B. (1975). An on-line transformation of EEG scalp potentials into orthogonal source derivations. *Electroencephalogr. Clin. Neurophysiol.* 39, 526–530. doi: 10.1016/0013-4694(75)90056-5

Hogervorst, M. A., Brouwer, A.-M., and van Erp, J. B. F. (2014). Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Front. Neurosci.* 8:322. doi: 10.3389/fnins.2014.00322

Hou, X., Liu, Y., Lim, W. L., Lan, Z., Sourina, O., Mueller-Wittig, W., et al. (2016). *CogniMeter: EEG-Based Brain States Monitoring*. Berlin; Heidelberg: Springer Berlin Heidelberg, 108–126.

International Civil Aviation Organization (2011). *Procedures for Air Navigation Services-Air Traffic Management, 14 Edn*. Montréal, QC: ICAO.

Jordan, C. (1992). *Experimental Study of the Effect of an Instantaneous Self Assessment Workload Recorder on Task Performance*. Technical Report DRA Technical Memorandum (CAD5) 92011, DRA Maritime Command Control Division.

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall.

Ke, Y., Qi, H., He, F., Liu, S., Zhao, X., Zhou, P., et al. (2014). An EEG-based mental workload estimator trained on working memory task can work well under simulated multi-attribute task. *Front. Hum. Neurosci.* 8:703. doi: 10.3389/fnhum.2014.00703

Kirwan, B., Evans, A., Donohoe, L., Kilner, A., Lamoureux, T., Atkinson, T., et al. (1997). "Human factors in the ATM system design life cycle," in *FAA/Eurocontrol ATM R&D Seminar* (Saclay).

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res. Rev.* 29, 169–195. doi: 10.1016/S0165-0173(98)00056-3

Kohlmorgen, J., Dornhege, G., Braun, M. L., Blankertz, B., Müller, K. R., Curio, G., et al. (2007). "Improving human performance in a real operating environment through real-time mental workload detection," in *Towards Brain-Computer Interfacing*, eds G. Dornhege, J. del R. Millán, T. Hinterberger, D. McFarland, and K. Müller (Cambridge, MA: MIT Press), 409–422.

Kompier, M. A. J., and Kristensen, T. S. (2001). "Organisational work stress interventions in a theoretical, methodological and practical context," in *Stress in the Workplace: Past, Present and Future*, ed J. Dunham (London: Whurr Publishers), 164–190.

Landsbergis, P. A., Cahill, J., and Schnall, P. (2003). The changing organisation of work and the safety and health of working people: a commentary. *J. Occupat. Environ. Med.* 45, 61–72. doi: 10.1097/00043764-200301000-00014

Lei, S., and Roetting, M. (2011). Influence of task combination on EEG spectrum modulation for driver workload estimation. *Hum. Fact.* 53, 168–179. doi: 10.1177/0018720811400601

Lin, C.-T., Ko, L.-W., Chung, I.-F., Huang, T.-Y., Chen, Y.-C., Jung, T.-P., et al. (2006). Adaptive EEG-based alertness estimation system by using ICA-based fuzzy neural networks. *IEEE Trans. Circuits Syst. I Reg. Pap.* 53, 2469–2476. doi: 10.1109/TCSI.2006.884408

Lohmann-Haislah, A. (2012). *Stressreport Deutschland 2012: Psychische Anforderungen, Ressourcen und Befinden*. Dortmund: Bundesanstalt für Arbeitsschutz und Arbeitsmedizin.

Lopez-Gordo, M., Sanchez-Morillo, D., and Valle, F. (2014). Dry EEG electrodes. *Sensors* 14, 12847–12870. doi: 10.3390/s140712847

Makeig, S., Bell, A. J., Jung, T.-P., and Sejnowski, T. J. (1996). "Independent component analysis of electroencephalographic data," in *Advances in Neural Information Processing Systems 8*, eds D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (Cambridge, MA: MIT Press), 145–151.

McEvoy, L. K., Pellouchoud, E., Smith, M. E., and Gevins, A. (2001). Neurophysiological signals of working memory in normal aging. *Cogn. Brain Res.* 11, 363–376. doi: 10.1016/S0926-6410(01)00009-X

Mecklinger, A., Kramer, A. F., and Strayer, D. L. (1992). Event related potentials and EEG components in a semantic memory search task. *Psychophysiology* 29, 104–119. doi: 10.1111/j.1469-8986.1992.tb02021.x

Mihajlovic, V., Grundlehner, B., Vullers, R., and Penders, J. (2015). Wearable, wireless EEG solutions in daily life applications: what are we missing? *IEEE J. Biomed. Health Inform.* 19, 6–21. doi: 10.1109/JBHI.2014.2328317

Mognon, A., Jovicich, J., Bruzzone, L., and Buiatti, M. (2011). ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology* 48, 229–240. doi: 10.1111/j.1469-8986.2010.01061.x

Niosh, N. (2002). *The Changing Organization of Work and the Safety and Health of Working People*. Technical Report 2002-116, National Institute for Occupational Safety and Health (NIOSH).

Omatu, S., Fujimura, M., and Kosaka, T. (2010). "Separation of noise and signals by independent component analysis," in *Fourth International Conference on Advanced Engineering Computing and Applications in Sciences 2010 (ADVCOMP 2010)*, ed W. Gentzsch (Florence: International Academy, Research, and Industry Association), 105–110.

Penaranda, B. N., and Baldwin, C. L. (2012). Temporal factors of EEG and artificial neural network classifiers of mental workload. *Proc. Hum. Fact. Ergon. Soc. Annu. Meeting* 56, 188–192. doi: 10.1177/1071181312561016

Pfurtscheller, G. (1997). EEG event-related desynchronization (ERD) and synchronization (ERS). *Electroencephalogr. Clin. Neurophysiol.* 103:26. doi: 10.1016/S0013-4694(97)88021-6

Pignat, J. M., Koval, O., Ville, D. V. D., Voloshynovskiy, S., Michel, C., and Pun, T. (2013). The impact of denoising on independent component analysis of functional magnetic resonance imaging data. *J. Neurosci. Methods* 213, 105–122. doi: 10.1016/j.jneumeth.2012.10.011

Puma, S., Matton, N., Paubel, P.-V., Raufaste, É., and El-Yagoubi, R. (2018). Using theta and alpha band power to assess cognitive workload in multitasking environments. *Int. J. Psychophysiol.* 123, 111–120. doi: 10.1016/j.ijpsycho.2017.10.004

Radüntz, T. (2017). Dual frequency head maps: a new method for indexing mental workload continuously during execution of cognitive tasks. *Front. Physiol.* 8:1019. doi: 10.3389/fphys.2017.01019

Radüntz, T. (2018). Signal quality evaluation of emerging eeg devices. *Front. Physiol.* 9:98. doi: 10.3389/fphys.2018.00098

Radüntz, T., and Meffert, B. (2019). User experience of 7 mobile electroencephalography devices: comparative study. *JMIR Mhealth Uhealth* 7:e14474. doi: 10.2196/14474

Radüntz, T., Scouten, J., Hochmuth, O., and Meffert, B. (2017). Automated EEG artifact elimination by applying machine learning algorithms to ICA-based features. *J. Neural Eng.* 14:046004. doi: 10.1088/1741-2552/aa69d1

Rosen, A., and Reiner, M. (2017). Right frontal gamma and beta band enhancement while solving a spatial puzzle with insight. *Int. J. Psychophysiol.* 122, 50–55. doi: 10.1016/j.ijpsycho.2016.09.008

Scerbo, M., Freeman, F., Mikulka, P., Parasuraman, R., Di Nocera, F., and Iii, L. (2001). *The Efficacy of Psychophysiological Measures for Implementing Adaptive Technology*. Hampton, VA: NASA - Langley Research Center.

Shou, G., Ding, L., and Dasari, D. (2012). Probing neural activations from continuous EEG in a real-world task: time-frequency independent component analysis. *J. Neurosci. Methods* 209, 22–34. doi: 10.1016/j.jneumeth.2012.05.022

Sterman, M. B., and Mann, C. A. (1995). Concepts and applications of EEG analysis in aviation performance evaluation. *Biol. Psychol.* 40, 115–130. doi: 10.1016/0301-0511(95)05101-5

Weiland, M. Z., Roberts, D. M., Fine, M. S., and Caywood, M. S. (2013). Real time research methods: Monitoring air traffic controller workload during simulation studies using electroencephalography (EEG). *Proc. Hum. Fact. Ergon. Soc. Annu. Meeting* 57, 1615–1619. doi: 10.1177/1541931213571359

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theor. Issues Ergon. Sci.* 3, 159–177. doi: 10.1080/14639220210123806

Wilson, G. F., and Russell, C. A. (2003a). Operator functional state classification using multiple psychophysiological features in an air traffic control task. *Hum. Fact.* 45, 381–389. doi: 10.1518/hfes.45.3.381.27252

Wilson, G. F., and Russell, C. A. (2003b). Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Hum. Fact.* 45, 635–643. doi: 10.1518/hfes.45.4.635.27088

Wilson, G. F., and Russell, C. A. (2007). Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. *Hum. Fact.* 49, 1005–1018. doi: 10.1518/001872007X249875

Winkler, I., Debener, S., Muller, K.-R., and Tangermann, M. (2015). "On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Milan: IEEE). doi: 10.1109/EMBC.2015.7319296

Xie, B., and Salvendy, G. (2000). Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments. *Work Stress* 14, 74–99. doi: 10.1080/026783700417249

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

SCIENTIFIC REPORTS
nature research

OPEN

The Effect of Planning, Strategy Learning, and Working Memory Capacity on Mental Workload

Thea Radüntz

In our modern society, planning and problem solving are crucial for handling a wide range of situations. Investigation of the experienced mental workload connected to planning, strategy learning, and working memory capacity is of particular interest for adjusting conditions according to the mental state of the individual. In our study, we examined 21 subjects during a planning and a working memory task. We applied the method of Dual Frequency Head Maps (DFHM) from the electroencephalogram for capturing mental workload objectively. We evaluated the DFHM-workload index and performance data during the learning and main phase of the planning task and linked the results to subjects' working memory capacity. The DFHM-workload index indicated that subjects with higher working memory capacity experienced a gradual decrease in mental workload during strategy learning of the planning task. However, the effect of learning on mental workload disappeared during the main phase.

Planning is a basic task in work and everyday life. In order to solve a problem, we firstly create a mental representation of the current situation and the goal state and plan the steps we need for transforming the initial state to the goal state[1]. Thereby, we generate multiple sequences of sub-goal states, rate their consequences, make decisions, and carry out actions, while continuously monitoring the outcome[2]. During planning, the working memory capacity plays an important role for maintaining and coordinating the sub-goal sequences[2–4]. Working memory defines the ability to temporally maintain information in mind and is linked not only to planning and problem solving but also to comprehension, reasoning, and learning[5]. Furthermore, working memory load is strongly connected to the experienced mental workload[6,7] that can be conceived as the amount of cognitive demands required for task solving related to the available cognitive resources[8–11].

Mental workload was often linked to mental health and human performance[12–18]. Objective registration and evaluation of mental workload is particular important in order to minimize errors and increase the safety of persons. Especially in our modern society, where planning and problem solving are crucial for handling a wide range of situations, the experienced workload as connected to planning, strategy learning, and working memory capacity is of particular interest. Understanding the interrelation between these constructs may contribute to adjust conditions, facilitate learning, enhance planning, and reduce mental workload.

A number of authors studied planning using the Tower of Hanoi (TOH) task and its connection to working memory[4,19–23] and found a connection between both[24,25]. Research on how planning and working memory relate to each other regarding their induced mental workload is rare. However, several researchers found that planning includes the interaction of working memory, inhibitory control, and cognitive flexibility and can be seen as a higher-order executive function that integrates core cognitive processes[26–28].

The study by Schiff and Vakil[29] investigated the connection between planning and learning. The authors employed the TOH task because they considered it to be particularly appropriate for the assessment of problem solving and learning of complex cognitive procedures. They stated that the learning phase starts with the first engagement with the task (i.e., subjects' baseline performance) and continues with rapid improvements during repeated practice within seconds to minutes. Study's findings emphasized a trade-off between younger and older children during the learning phase that became evident through faster speed and greater accuracy for the older ones. Schiff and Vakil[29] argued that there exists only one further study by Beaunieux *et al.*[30] that examined learning effects by means of the TOH. Aside from this, working memory is needed for concept formation and for controlling processes as well as remember strategies that are all important for learning[5]. Several studies suggested that learning can be facilitated by increased working memory capacity[31–37]. Thus, the relation between the

Federal Institute for Occupational Safety and Health, Work and Health, Mental Health and Cognitive Capacity, Berlin, 10317, Germany. e-mail: raduentz.thea@baua.bund.de

amount of available cognitive resources and cognitive demands required for task solving during learning should be reflected accordingly by registration of mental workload. A study that connects and investigates these aspects is not present yet.

Research also indicated a quick saturation after a fast learning effect[29,38,39]. Specifically, after a short learning phase the performance became stable for consecutive trials within a session[38]. Despite that, the performance might continue to improve again on subsequent daily sessions[38]. The time course of learning follows a curve that gradually reaches an asymptote but after intense practice and rehearsal the learned skill could become automatic[40]. This trend of fast improvement followed by a floor effect of performance can be observed also in the figures of the TOH study by Schiff and Vakil[29] for younger as well as older children. Human performance and mental workload were often linked to each other[17,18] and their relation was frequently outlined by the Yerkes-Dodson curve[41,42]. Consequently, a quick saturation in performance after a fast learning effect, should be also prevalent in the registration of mental workload.

As far as we know there exists only one study related to mental workload and planning. Hardy and Wright[43] manipulated the difficulty of the TOH task and assessed the workload using the NASA-TLX questionnaire[44] as a subjective method for workload registration. Thereby, workload ratings increased with increasing TOH difficulty and individual performance on the TOH correlated with the subjective ratings. The authors suggested that mental workload did not only reflect task's cognitive demands but also the cognitive abilities of the performer. That means that although subjects could reveal similar task performance, they might experience different levels of workload. Hardy and Wright[43] stated that measuring workload during cognitive tests provided additional information about the cognitive state of the subject and captured individual differences.

However, the assessment of workload using subjective questionnaire methods has a number of drawbacks. Subjective registration of mental workload is only possible in retrospect and the questionnaire method might alter subject's mental state by imposing additional demands. An objective and reliable method for measuring instantaneous mental workload continuously over time would be more beneficial.

Over the past 50 years, different physiological parameters (e.g., heart rate and derived parameters, electrodermal activity, body temperature, etc.) have been evaluated for their validity regarding continuous mental workload registration. In last century's 90 s, the ability of the electroencephalogram (EEG) for registering mental workload was evaluated and served as a starting point for the use of the EEG in applied research. Basically, changes in the alpha-frequency (8–12 Hz) and theta-frequency (4–8 Hz) band powers related to mental workload have been confirmed many times. Thereby, the majority of workload studies dealt with the analysis of the EEG during cognitive tasks related to working memory and executive control[45–49]. In a review article, Borghini *et al.*[50] provided a detailed overview of the measurement of neurophysiological signals for the determination of mental workload and confirmed essentially the known relations. In recent years, classifiers were increasingly used for the separation of workload levels. The feature vectors derived from the EEG revealed varying complexity and extent, and frequency bands were taken differently into account. The used EEG parameters were, for example, the amplitude of the EEG, spectral power of different frequency bands and different EEG channels[7,51–55]. The focus was on frontal, parietal, and occipital EEG channels according to previous findings. Independent component analysis (ICA) was used to determine specific reactions of spatio-temporal different sources[56] and allowed the successful detection and elimination of artifacts[57–59].

Nevertheless, different cognitive strategies in task solving, both intra- and inter-individually, can influence the classification results of mental workload. Additionally, the question arises whether machine learning algorithms provide reliable and reproducible results over time. In particular, the need for appropriate retraining of the classifier regarding subjects and tasks poses additional demands for the investigation of the interrelations between planning, strategy learning, working memory capacity, and mental workload. To the best of our knowledge, there is no other study currently available that investigated the interactions of working memory, learning effects during planning, and objective mental workload registration using the EEG.

In our prior work we developed a mental-workload classifier that does not need retraining, neither for new subjects nor for new tasks[60]. In a laboratory study conducted with 54 subjects which executed well-established cognitive tasks, we developed the so-called Dual Frequency Head Maps (DFHM). These head maps consist of personalized spectral features and their spatial occurrence (i.e., frontal theta-band and parietal alpha-band powers). Support vector machines are used for classification in three classes: low, moderate, or high workload. Under laboratory conditions, we successfully proved the DFHM method as universally applicable for mental-workload indexing.

In the current study, we applied the DFHM method for capturing mental workload objectively during a planning and a working memory task. We employed the TOH as a planning task and the automated orientation span (AOSPAN) task as a working memory task. The aim of our study was the investigation of the effect of planning, strategy learning, and working memory capacity on mental workload. In a first step, we aimed to show that a higher-level executive function like planning involving several core cognitive processes[26–28] imposes a higher mental workload than a working memory task as it binds more cognitive resources. Next, we investigated interrelations between planning, strategy learning, working memory capacity, and mental workload according to the last two hypotheses.

1. Execution of a planning task induces higher mental workload compared to a working memory task.
2. A higher working memory capacity contributes to a better strategy learning and thus to a gradual decrease in mental workload during the learning phase of the planning task.
3. After the learning phase, the effect of strategy learning on mental workload disappears during increasing task load.

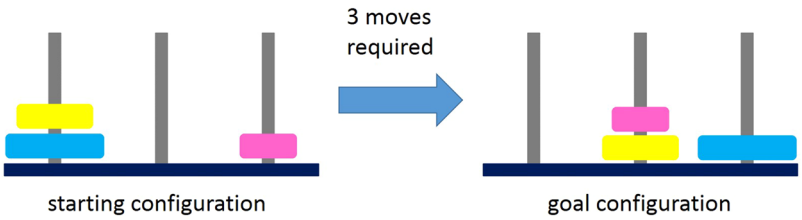


Figure 1. Computerized version of Tower of Hanoi. Subjects were required to transform the starting configuration into the goal configuration by three moves.

Methods

Procedure, Tasks, and Subjects. For our investigation we employed the TOH and AOSPAN tasks. Their implementation was realized with the E-Prime application suite. All subjects executed both tasks in counterbalanced order.

The TOH task consists of three pegs with discs of graduated size. Subjects were asked to transform the starting configuration into a given goal configuration (Fig. 1) in as few moves as possible. For this, they had to select a top disc from the source peg and place it to a destination peg. They were allowed to move only one disc at a time and they were not allowed to place big discs on smaller ones. The experiment started with a small instruction procedure where the TOH task was explained to the subjects. For familiarizing themselves with the clicking procedure during the task, subjects were asked to execute three trials with 1, 2, and 3 moves required to reach the goal configuration. Thereafter, the main experiment started including a learning phase and a main phase. The learning phase consisted of 3 trials with 3 discs each and 5, 6, and 7 moves required to reach the goal state. The main phase consisted of 3 trials with 4 discs and 7, 11, and 15 moves. In order to reach the goal-state configurations with the least-possible moves, subjects were instructed to plan their actions before starting. The number of least-required moves was given to them before each trial. If a move was not optimal and would result in a greater number of moves, they got an error message and had to start the trial again. There was no time limit set, neither for the planning time nor for task solution in general, for avoiding the tendency of a speed and accuracy trade-off. Furthermore, subjects should make full usage of the time before their first move, which was used later for performance evaluation of planning time, instead of planning during the movements.

The AOSPAN task was administered as a working memory task in the version developed by Unsworth *et al.*[61]. It was translated in German and adapted accordingly. Subjects were asked to memorize letters in the order presented while simultaneously solving math problems. The math problems required to click as soon as subjects knew the answer. After the click a number was presented and subjects had to judge if it was the right answer to the problem. Then a letter to be memorized was shown. At the end a recall slide was presented asking them to select the letters shown in the correct order. Finally, subjects got feedback about both their memory and math performance. Furthermore, the subjects were instructed to keep the percentage number indicating their math performance above 85%. The AOSPAN training took place directly before the actual task as described in Unsworth *et al.*[61]. The math practice of the task aimed to calculate for each person how long they needed to solve the math problems. Each individual's mean (plus 2.5 SD) was used during the main AOSPAN task as a time limit for the math operations in order to account for individual differences. According to Unsworth *et al.*[61], the time limit serves to prevent participants from rehearsing the letters when they should be solving the operations.

The participating subjects needed about 25 min to complete both tasks. Performance evaluation for the TOH task was done by analysis of individual error rates and planning time until their first move. The working memory capacity of the subjects reflected by the AOSPAN task was calculated by means of the sum of correctly recalled letters from only the sets in which all characters were recalled in correct serial order. Similar to Unsworth *et al.*[61], we refer to it as absolute score.

We examined 21 subjects in the age between 22 and 64 years (2 female, 19 male, mean age 38 $\pm$ 11). All subjects had a background in science or engineering associate education. All of the investigations acquired were approved by the local review board of our institution and complied with the tenets of the Declaration of Helsinki. All procedures were carried out with the adequate understanding and written consent of the subjects.

EEG and DFHM-Workload index. Biosignal processing and all calculations were done with MATLAB.

For EEG registration we used g.tec's g.LADYbird/g.Nautilus system with 25 active electrodes placed at positions according to the 10–20 system (Fig. 2). Registration was carried out with a sample rate of 500 Hz and with reference to electrode Cz. For signal recording we used g.tec's Matlab interface.

After recording, the EEG was filtered with a bandpass filter (order 100) between 0.5 and 40 Hz. Independent component analysis (ICA, Infomax algorithm[62]) for artifact rejection was applied to the signal. In order to increase topographical localization, we applied a simple Hjorth-style surface Laplacian filter using 8 neighbours[63]. This spatial high-pass filter was aimed to attenuate large-scale scalp signals and amplify localized signals.

The artefact-free EEG was transformed to average reference and cut into segments of 1 s length, overlapping by 0.5 s. By means of Fast Fourier Transformation (FFT) we computed the workload relevant frequency bands (theta: 4–8 Hz, alpha: 8–12 Hz) over the segments and generated the DFHM as outlined in the article by Radüntz[60]. In brief, we applied a theta-bandpass filter to the signals of the frontal electrodes and an alpha-bandpass filter to the signals of the parietal electrodes and calculated for each participant, each electrode, and each segment the z-scores of theta and alpha band power. The compilation of the z-scores of the theta band power from the frontal

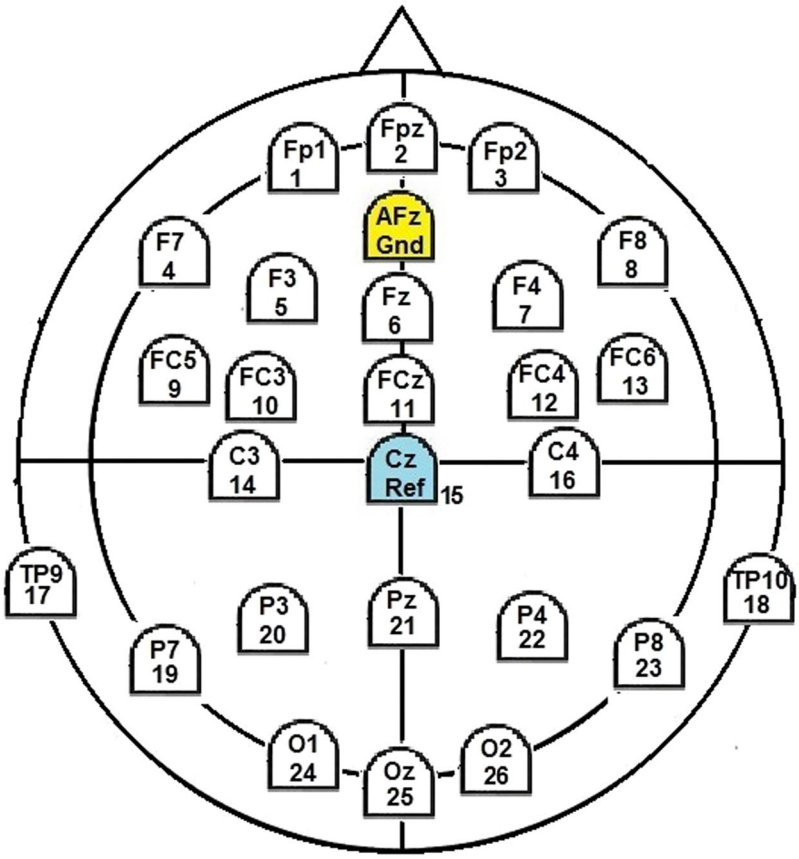


Figure 2. EEG layout used.

electrodes and alpha band power from the parietal electrodes constituted the DFHM for each EEG segment. The individual mean and standard deviation for z-score calculation were obtained from subject's segments of the first minute of EEG recordings. These consisted not only of the two tasks relevant for this article but also of six rest measurements and eight different workplace tasks familiar to the subjects. They were conducted during a following two-day experiment and are not subject of this article.

We used the already trained SVM classifiers from the laboratory study[60] to classify the DFHM of each subject from the tasks' segments. Every 0.5 s we obtained a value determining if the segment belongs to low, moderate, or high workload. We applied a moving-average time window of 6 s and adjusted the result in order to gain a DFHM-workload index as percentage value between 0 (all DFHM classified as low) and 100 (all DFHM classified as high).

Statistical analysis. For evaluating our first hypothesis confirming the expected higher mental workload during the the planning task, we calculated the DFHM-index average over the TOH and AOSPAN tasks. The Shapiro-Wilk test did not show normal distribution for the differences of the DFHM-index averages between both tasks. Thus, a Wilcoxon signed-rank test was calculated.

For investigating the effect of working memory capacity on mental workload during strategy learning of a planning task (hypothesis 2), we employed the DFHM-index averages of the three TOH trials of the learning phase. The Shapiro-Wilk test showed a normal distribution for the three DFHM-index averages. Thus, we carried out an analysis of variance (ANOVA) with the items' mean DFHM index as dependent variable. We utilized a repeated-measures design with one within-subject factor for the number of required moves (three levels: 5, 6, and 7 moves) and one between-subject factor for the working memory capacity (two levels). The latter was calculated using the median of the absolute score of the AOSPAN task. Subjects with an absolute score below the median of 43 were classified as low working memory capacity subjects ($n = 10$), the remaining as subjects with high mental workload capacity ($n = 11$). General differences between the levels were examined and tested with a post-hoc test (Bonferroni corrected). Additionally, we evaluated subjects' planning times and the number of errors (i.e. number of restarts) for each TOH trial. The Shapiro-Wilk test did not show normal distribution, neither for the planning time nor for the number of errors. For achieving a normal distribution for the further analysis, we computed the logarithm of the planning time. Thus, we were able to proceed in the same way as described above and conduct a repeated-measures mixed ANOVA with one within-subject and one between-subject factor. Computation of the logarithm of the number of errors did not yield normal distribution. Hence, statistical analysis of the number of errors was conducted via non-parametric Friedman test of differences among the repeated measures. Dunn-Bonferroni post-hoc tests were calculated for the examination of the differences between the levels.

	DFHM-workload index	Planning time [s]	Errors
Condition	**Mean ± SD, median [min, max]**	**Mean ± SD, median [min, max]**	**Mean ± SD, median [min, max]**
AOSPAN, whole task[a]	57.5 ± 6.0, 56.6 [48.4, 67.5]	–	–
TOH, whole task[a]	62.5 ± 7.4, 63.6 [42.5, 73.9]	–	–
TOH learning, 5 moves			
Lower WM capacity[b]	63.0 ± 6.9, 64.8 [52.2, 72.0]	22.5 ± 16.2, 20.2 [5.1, 58.5]	1.2 ± 2.2, 0 [0, 7]
Higher WM capacity[c]	65.4 ± 10.3, 65.0 [43.8, 81.5]	20.5 ± 12.2, 17.1 [5.3, 42.5]	0.9 ± 1.2, 1 [0, 4]
TOH learning, 6 moves			
Lower WM capacity[b]	62.9 ± 7.8, 64.8 [47.9, 70.4]	14.1 ± 9.6, 12.1 [4.9, 37.5]	0.6 ± 1.6, 0 [0, 5]
Higher WM capacity[c]	61.8 ± 11.1, 62.9 [35.3, 75.5]	16.2 ± 11.8, 10.3 [5.1, 38.0]	0.5 ± 0.9, 0 [0, 3]
TOH learning, 7 moves			
Lower WM capacity[b]	64.9 ± 8.1, 66.4 [51.8, 77.0]	20.3 ± 13.4, 17.1 [6.9, 45.4]	1 ± 1.3, 0.5 [0, 4]
Higher WM capacity[c]	60.7 ± 9.0, 59.8 [48.4, 79.0]	15.7 ± 11.2, 11.8 [4.5, 36.0]	0.6 ± 0.7, 1 [0, 2]
TOH main, 7 moves			
Lower WM capacity[b]	63.9 ± 8.0, 66.0 [48.1, 76.6]	10.9 ± 7.8, 7.8 [4.5, 26.4]	0.2 ± 0.6, 0 [0, 2]
Higher WM capacity[c]	60.2 ± 10.7, 63.4 [40.0, 73.2]	15.3 ± 9.7, 10.5 [5.2, 35.0]	0.2 ± 0.4, 0 [0, 1]
TOH main, 11 moves			
Lower WM capacity[b]	63.4 ± 6.1, 64.3 [53.1, 70.4]	20.1 ± 17.7, 15.4 [6.8, 64.6]	1.1 ± 1.4, 1 [0, 4]
Higher WM capacity[c]	63.1 ± 11.4, 62.1 [37.6, 78.7]	16.5 ± 13.4, 9.0 [2.6, 38.2]	0.6 ± 0.8, 0 [0, 2]
TOH main, 15 moves			
Lower WM capacity[b]	64.3 ± 5.8, 65.3 [53.8, 71.4]	18.8 ± 10.4, 18.2 [8.9, 45.5]	2.6 ± 4.2, 1 [0, 13]
Higher WM capacity[c]	64.4 ± 10.0, 66.7 [42.7, 76.0]	19.4 ± 15.7, 12.8 [6.5, 56.3]	1 ± 1.6, 0 [0, 5]

Table 1. Descriptive statistics of the dependent variables related to research hypotheses' conditions (WM: working memory). *Note.* [a]All subjects: N = 21, [b]Subjects with lower WM capacity: N = 10, [c]Subjects with higher WM capacity: N = 11.

Finally, we addressed the issue of mental workload related to planning after the learning phase (hypothesis 3). We employed the DFHM-index averages, planning times, and number of errors of the three TOH trials during the main phase. The Shapiro-Wilk test showed similar results for all variables as during the learning phase. We carried out two repeated-measures mixed ANOVA with one within-subject and one between-subject factor, one for the DFHM index and one for the logarithm of the planning time. A non-parametric Friedman test of differences was conducted for the number of errors among the repeated measures. Dunn-Bonferroni post-hoc tests were calculated for the examination of the differences between the levels.

Statistical calculations were conducted using SPSS and the significance threshold was set at 5%.

Results

Planning task causes higher mental workload than working memory task. The Wilcoxon signed-rank test indicated significant mental workload differences between the TOH and AOSPAN tasks ($T = 26$, $z = -3.11$, $p = 0.002$, $r = 0.48$). The mental workload assessed by the DFHM-workload index from the EEG was higher for the TOH than for the AOSPAN task. Descriptive statistics are presented in Table 1 and Fig. 3 shows the results.

Higher working memory capacity contributes to workload decrease during strategy learning of planning. The mixed ANOVA yielded a significant interaction effect of requested moves and working memory capacity on mental workload ($F(2, 38) = 3.62$, $p = 0.036$, $\eta^2 = 0.159$). For subjects with higher working-memory capacity the DFHM-workload decreased during the learning phase. Post-hoc analysis indicated that workload decreased significantly from the initial to the second ($p = 0.031$) and third trial ($p = 0.025$). For subjects with lower working-memory capacity, we were not able to obtain any significant workload differences during the learning phase. Evaluation of planning time and errors did not reveal any significant effects for the number of requested moves or subjects' working memory capacity during the learning phase. Descriptive statistics are presented in Table 1 and Fig. 4 presents the results.

Learning effect on mental workload disappears after the learning phase. During the main phase, no significant learning effect could be obtained. This applied for mental workload as well as for planning time where mixed ANOVA calculations showed no significant effects of the number of requested moves or subjects' working memory capacity. The non-parametric Friedman test revealed a general significant change in the number of errors for the lower working memory capacity subjects ($\chi^2 = 8.960$, $df = 2$, $n = 10$, $p < 0.011$). Nevertheless, subsequently conducted post-hoc tests did not reveal significant differences between the levels. For the higher working memory capacity subjects this effect was not prominent at all. Descriptive statistics are presented in Table 1 and Fig. 5 illustrates the results.

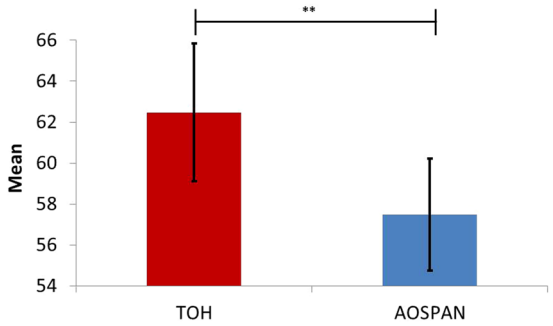


Figure 3. Mean DFHM-workload index during TOH and AOSPAN tasks (Wilcoxon signed-rank test differences: $^{**}0.001 < p \leq 0.01$; error bars indicate 95% confidence interval).

Discussion

In our study, we investigated the effect of planning, strategy learning, and working memory capacity on mental workload. For assessing mental workload, we used the DFHM method that was previously developed in a laboratory setting and is based on the EEG. In the current study, 21 subjects participated and completed the TOH and AOSPAN tasks in randomized order. We registered the EEG and computed the DFHM-workload index for each subject and task. We did not retrain the classifiers neither for the new tasks nor for the new subjects.

The DFHM-workload index was significantly higher for the TOH than for AOSPAN task as stated by hypothesis 1. This indicated that planning imposed higher mental workload suggesting that more cognitive resources were required during planning than working memory task. The result was consistent with literature that stated that planning is a higher-order executive function that integrates core cognitive processes such as working memory, inhibitory control, and cognitive flexibility[26–28]. Although attentive readers could argue that the time limit set for the math operations during the AOSPAN task might result in time pressure and increase mental workload, our results did not support this assumption.

More insight regarding intra-individual differences linked to strategy learning and mental workload during planning was gained from subject clustering by means of the absolute score from the AOSPAN task as an indicator for subjects' working memory capacity. During the learning phase of the TOH task, we were able to obtain a significant interaction effect between task load and working memory capacity on mental workload. Thereby, mental workload of subjects with higher working memory capacity significantly decreased while the workload of subjects with lower working memory capacity did not yield significant changes. The effect was particularly prominent for the mental workload assessed by the EEG whereas the number of errors and planning time showed only a weak tendency in that direction. This fits well the assumption by Hardy and Wright[43] that mental workload reflects the cognitive abilities of the performer, captures individual differences, and reveals additional information about the cognitive state although task performance might be similar. We concluded that a higher working memory capacity contributes to workload decrease during strategy learning of planning as suggested by hypothesis 2. Nevertheless, learning is traditionally associated with a change in behavior[64] and one could ask if a reduction in mental workload can indicate a learning process when there is no such change. According to the definitions of different authors[8–11], mental workload reflects the amount of cognitive resources required for task solving. In our experiments, subjects with higher working memory capacity needed less cognitive resources for maintaining their performance although the number of required moves gradually increased during the learning phase. Consequently, we suggested that this result indicated an initial learning process on neurological level that might produce behavioral changes after longer practice. Considering the obtained tendency of performance enhancement, this assumption seems rational. However, further studies should allow subjects to perform the same version of the task more times for providing statistical-significant evidence. A possible explanation that performance changes did not reach the significance level might be also related to the higher educational background of our subjects. This might have impacted the performance by a floor effect as well. Finally, we want to call attention to a study by Huang *et al.*[65] with supporting results for our assumption. The research was concerned with driving learning. The authors found that later stages of motor learning increased metabolic efficiency but did not reveal any gains in performance.

As task load of the planning task increased during the main phase of the TOH, the learning effect disappeared and mental workload increased regardless of subjects' working memory capacity. The DFHM-workload index of both subject clusters converged at the most demanding trial. Conforming to hypothesis 3 results indicated a quick saturation after the short learning phase. This was particularly true for subjects with higher working memory capacity that have previously experienced a fast learning effect. Even though we were able to detect a tendency to more errors for the subjects with lower working memory capacity, the pairwise comparisons between the levels did not become significant for none of our variables. The subjects with lower working memory capacity did not seem to have learned the task at all, since at no point did the DFHM-workload index display refinement nor did performance improve. In addition, in the main phase of the experiment, the performance of the low working memory capacity group tended to reduce with no apparent change in workload. In other words, although subjects invested the same amount of cognitive resources their performance got worst with increasing task difficulty. All facts together support our previous suggestion that mental workload indicates an initial learning process on neurological level that may result in behavioral changes during the main practice.

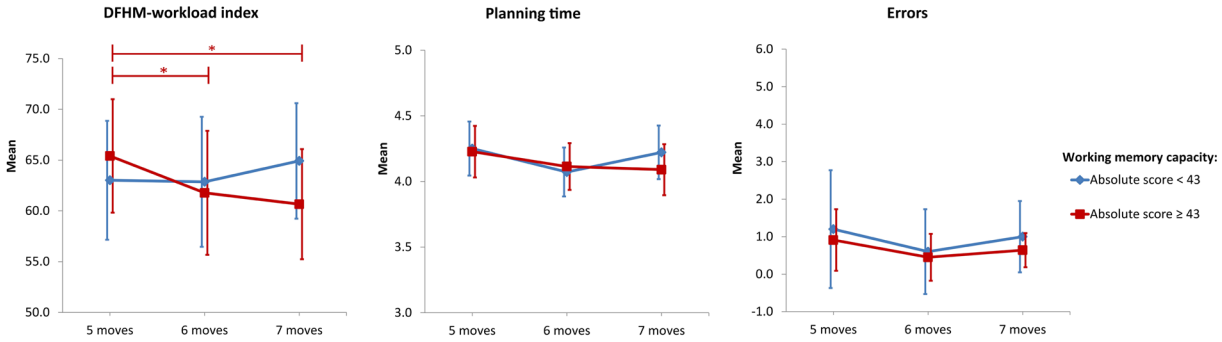


Figure 4. Mean values of DFHM-workload index (left), logarithm of planning time (middle), and errors (right) during the learning phase of the TOH task for subjects with lower (blue) and higher (red) working memory capacity as indicated by the absolute score of the AOSPAN task (Bonferroni corrected post-hoc tests: *$0.01 < p \leq 0.05$; error bars indicate 95% confidence interval).

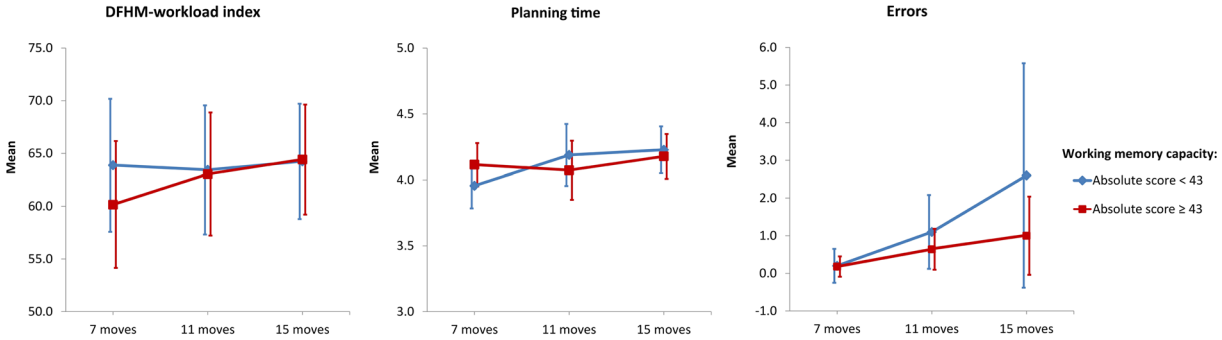


Figure 5. Mean values of DFHM-workload index (left), logarithm of planning time (middle), and errors (right) during the main phase of the TOH task for subjects with lower (blue) and higher (red) working memory capacity as indicated by the absolute score of the AOSPAN task (error bars indicate 95% confidence interval).

A limitation of our study was our small sample set. Future studies should involve more females, subjects with different educational levels, and also older participants. In our study, the educational background of our subjects was in science or engineering and equally high among them. Affinity with the underlying tasks might have affected subjects' performance and mental workload. The investigation of older subjects in connection to learning and mental workload is particularly relevant and meets the evolving needs and expectations of the demographic change of our society and the challenge of life-long learning. An objective method for continuous mental workload registration can offer a way for understanding procedural learning, enhancing skill acquisition, and identifying possible risks.

To conclude, our study was concerned with the neuronal registration of mental workload as connected to planning, strategy learning, and working memory capacity. The topic is of particular interest because of the importance of these constructs for handling a wide range of situations in our digitized world. Understanding the interrelation among them may contribute to adjust conditions, facilitate learning, enhance planning, and reduce workload in accordance to the cognitive abilities of the individual. To the best of our knowledge, there is no other study that investigated planning and mental workload by means of the EEG. We demonstrated the capability of the DFHM index from the EEG to successfully register mental workload and suggest the DFHM method as a useful tool for further studies. In our future research, we aim at employing the DFHM index for the investigation of mental workload related issues of the modern society.

Data availability

The conducted data used to support the findings of this study are restricted by the ethics committee of the Federal Institute for Occupational Safety and Health in order to protect subjects privacy according to data-protection regulations. Data can be made available from the corresponding author upon request and after approval of the legal department for researchers who meet the criteria for access to confidential data.

Received: 8 October 2019; Accepted: 7 April 2020;
Published online: 27 April 2020

References

1. Sternberg, R. J. & Ben-Zeev, T. *Complex Cognition: The Psychology of Human Thought* (OXFORD UNIV PR, 2001).
2. Carlin, D. *et al*. Planning impairments in frontal lobe dementia and frontal lobe lesion patients. *Neuropsychol.* **38**, 655–665, https://doi.org/10.1016/s0028-3932(99)00102-5 (2000).
3. Goel, V. & Grafman, J. Are the frontal lobes implicated in "planning" functions? interpreting data from the tower of hanoi. *Neuropsychol.* **33**, 623–642, https://doi.org/10.1016/0028-3932(95)90866-p (1995).
4. Goela, V., Pullara, D. & Grafman, J. A computational model of frontal lobe dysfunction: working memory and the tower of hanoi task. *Cognitive Science* **25**, 287–313, https://doi.org/10.1207/s15516709cog2502_4 (2001).
5. Cowan, N. Working memory underpins cognitive development, learning, and education. *Educational Psychology Review* **26**, 197–223, https://doi.org/10.1007/s10648-013-9246-y (2013).
6. Brouwer, A. M. *et al*. Estimating workload using EEG spectral power and ERPs in the n-back task. *Journal of Neural Engineering* **9**, 045008, https://doi.org/10.1088/1741-2560/9/4/045008 (2012).
7. Ke, Y. *et al*. An EEG-based mental workload estimator trained on working memory task can work well under simulated multi-attribute task. *Frontiers in Human Neuroscience* **8**, 703, https://doi.org/10.3389/fnhum.2014.00703 (2014).
8. Eggemeier, F., Wilson, G. F., Kramer, A. F. & Damos, D. L. *Multiple-task performance*, chap. Workload assessment in multi-task environments, 207–216 (Taylor & Francis, 1991).
9. Kahneman, D. *Attention and Effort* (Prentice-Hall, Englewood Cliffs, 1973).
10. Wickens, C. D. Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science* **3**, 159–177, https://doi.org/10.1080/14639220210123806 (2002).
11. Xie, B. & Salvendy, G. Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments. *Work & Stress* **14**(1), 74–99 (2000).
12. Zoer, I., Ruitenburg, M. M., Botje, D., Frings-Dresen, M. H. W. & Sluiter, J. K. The associations between psychosocial workload and mental health complaints in different age groups. *Ergonomics* **54**, 943–952, https://doi.org/10.1080/00140139.2011.606920 PMID: 21973005 (2011).
13. Klonowicz, T. Mental workload and health: A latent threat. *International Journal of Occupational Safety and Ergonomics* **1**, 130–135, https://doi.org/10.1080/10803548.1995.11076309 PMID: 10603543 (1995).
14. Kompier, M. A. J. & Kristensen, T. S. Organisational work stress interventions in a theoretical, methodological and practical context. In Dunham, J. (ed.) *Stress in the Workplace: Past, Present and Future*, 164–190 (Whurr Publishers, London, 2001).
15. Landsbergis, P. A., Cahill, J. & Schnall, P. The changing organisation of work and the safety and health of working people: A commentary. *Journal of Occupational Environmental Medicine* **45**, 61–72, https://doi.org/10.1097/00043764-200301000-00014 (2003).
16. NIOSH, N. The changing organization of work and the safety and health of working people. Tech. Rep. 2002–116, National Institute for Occupational Safety and Health (NIOSH) (2002).
17. Parasuraman, R., Molloy, R. & Singh, I. L. Performance consequences of automation induced complacency. *International Journal of Aviation Psychology* **3**, 1–23 (1993).
18. Sträter, O. Warum passieren menschliche fehler und was kann man dagegen tun? In *Forum Prävention* (AUVA - Allgemeine Unfallversicherungsanstalt, Wien, 2001).
19. Lehto, J. Are executive function tests dependent on working memory capacity? *The Quarterly Journal of Experimental Psychology Section A* **49**, 29–50, https://doi.org/10.1080/713755616 (1996).
20. Colom, R., Rubio, V. J., Shih, P. C. & Santacreu, J. Fluid intelligence, working memory and executive functioning. *Psicothema* **18**, 816–821 (2006).
21. Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P. & Hegarty, M. How are visuospatial working memory, executive functioning, and spatial abilities related? a latent-variable analysis. *Journal of Experimental Psychology: General* **130**, 621–640, https://doi.org/10.1037/0096-3445.130.4.621 (2001).
22. Numminen, H., Lehto, J. E. & Ruoppila, I. Tower of hanoi and working memory in adult persons with intellectual disability. *Research in Developmental Disabilities* **22**, 373–387, https://doi.org/10.1016/s0891-4222(01)00078-6 (2001).
23. Zook, N. A., Davalos, D. B., DeLosh, E. L. & Davis, H. P. Working memory, inhibition, and fluid intelligence as predictors of performance on tower of hanoi and london tasks. *Brain and Cognition* **56**, 286–292, https://doi.org/10.1016/j.bandc.2004.07.003 (2004).
24. Chan, R. C. K., Wang, Y. N., Cao, X. Y. & Chen, E. Y. H. Contribution of working memory components to the performance of the tower of hanoi in schizophrenia. *East Asian archives of psychiatry: official journal of the Hong Kong College of Psychiatrists = Dong Ya jing shen ke xue zhi: Xianggang jing shen ke yi xue yuan qi kan* **20**, 69–75 (2010).
25. Handley, S. J., Capon, A., Copp, C. & Harper, C. Conditional reasoning and the tower of hanoi: the role of spatial and verbal working memory. *British journal of psychology (London, England: 1953)* **93**, 501–518 (2002).
26. Ávila, R. T. *et al*. Working memory and cognitive flexibility mediates visuoconstructional abilities in older adults with heterogeneous cognitive ability. *Journal of the International Neuropsychological Society* **21**, 392–398, https://doi.org/10.1017/s135561771500034x (2015).
27. Diamond, A. Executive functions. *Annual Review of Psychology* **64**, 135–168, https://doi.org/10.1146/annurev-psych-113011-143750 (2013).
28. Miyake, A. *et al*. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology* **41**, 49–100, https://doi.org/10.1006/cogp.1999.0734 Last accessed on 2014-03-17 (2000).
29. Schiff, R. & Vakil, E. Age differences in cognitive skill learning, retention and transfer: The case of the tower of hanoi puzzle. *Learning and Individual Differences* **39**, 164–171, https://doi.org/10.1016/j.lindif.2015.03.010 (2015).
30. Beaunieux, H. *et al*. Which processes are involved in cognitive procedural learning? *Memory* **14**, 521–539, https://doi.org/10.1080/09658210500477766 (2006).
31. Hulme, C. *Working Memory and Severe Learning Difficulties (PLE: Memory)* (Psychology Press, 2014).
32. Brandenburg, J. *et al*. Working memory in children with learning disabilities in reading versus spelling. *Journal of Learning Disabilities* **48**, 622–634, https://doi.org/10.1177/0022219414521665 (2014).
33. Alloway, T. P. Working memory, but not IQ, predicts subsequent learning in children with learning difficulties. *European Journal of Psychological Assessment* **25**, 92–98, https://doi.org/10.1027/1015-5759.25.2.92 (2009).
34. Wagner, A. D. Working memory contributions to human learning and remembering. *Neuron* **22**, 19–22, https://doi.org/10.1016/s0896-6273(00)80674-1 (1999).
35. Swanson, H. L. Working memory in learning disability subgroups. *Journal of Experimental Child Psychology* **56**, 87–114, https://doi.org/10.1006/jecp.1993.1027 (1993).
36. Woltz, D. J. An investigation of the role of working memory in procedural skill acquisition. *Journal of Experimental Psychology: General* **117**, 319–331, https://doi.org/10.1037/0096-3445.117.3.319 (1988).
37. Baddeley, A. D. & Hitch, G. Working memory. In *Psychology of Learning and Motivation*, 47–89, https://doi.org/10.1016/s0079-7421(08)60452-1 (Elsevier, 1974).
38. Karni, A. & Sagi, D. The time course of learning a visual skill. *Nature* **365**, 250–252, https://doi.org/10.1038/365250a0 (1993).
39. Callan, D. E. *et al*. Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language phonetic contrast. *NeuroImage* **19**, 113–124, https://doi.org/10.1016/S1053-8119(03)00020-X (2003).

40. Stickgold, R. & Walker, M. Memory consolidation and reconsolidation: what is the role of sleep? *Trends in Neurosciences* **28**, 408–415, https://doi.org/10.1016/j.tins.2005.06.004 (2005).
41. Yerkes, R. M. & Dodson, J. D. The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology* **18**, 459–482 Last accessed on 2011-11-03 (1908).
42. de Waard, D. *The measurement of drivers' mental workload*. Ph.D. thesis, University of Groningen, Traffic Research Centre, Haren, Netherlands (1996).
43. Hardy, D. J. & Wright, M. J. Assessing workload in neuropsychology: An illustration with the tower of hanoi test. *Journal of Clinical and Experimental Neuropsychology* **40**, 1022–1029, https://doi.org/10.1080/13803395.2018.1473343 (2018).
44. Hart, S. G. & Staveland, L. E. Development of the NASA TLX: results of empirical and theoretical research. In Hancock, P. & Meshkati, N. (eds.) *Human Mental Workload*, 139–183 (North Holland, Amsterdam., 1988).
45. Klimesch, W. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews* **29**, 169–195 (1999).
46. Gevins, A. *et al*. Monitoring working memory load during computer-based tasks with EEG pattern recognition methods. *Human Factors: The Journal of the Human Factors and Ergonomics Society* **40**, 79–91, https://doi.org/10.1518/001872098779480578 (1998).
47. Gevins, A. & Smith, M. E. Neurophysiological measures of working memory and individual differences in cognitive ability and cognitive style. *Cerebral Cortex* **10**, 829–839, https://doi.org/10.1093/cercor/10.9.829 Last accessed on 2014-02-18 (2000).
48. Pfurtscheller, G. EEG event-related desynchronization (ERD) and synchronization (ERS). *Electroencephalography and Clinical Neurophysiology* **103**, 26 (1997).
49. Sterman, M. B. & Mann, C. A. Concepts and applications of EEG analysis in aviation performance evaluation. *Biological Psychology* **40**, 115–130, https://doi.org/10.1016/0301-0511(95)05101-5 EEG in Basic and Applied Settings (1995).
50. Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D. & Babiloni, F. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews* **44**, 58–75, https://doi.org/10.1016/j.neubiorev.2012.10.003 Applied Neuroscience: Models, methods, theories, reviews. A Society of Applied Neuroscience (SAN) special issue (2014).
51. Baldwin, C. L. & Penaranda, B. N. Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification. *NeuroImage* **59**, 48–56, https://doi.org/10.1016/j.neuroimage.2011.07.047 Neuroergonomics: The human brain in action and at work (2012).
52. Kohlmorgen, J. *et al*. Improving human performance in a real operating environment through real-time mental workload detection. In Dornhege, G., del R. Millán, J., Hinterberger, T., McFarland, D. & Müller, K. (eds.) *Towards Brain-Computer Interfacing*, 409–422 (MIT Press, Cambridge, 2007).
53. Lin, C.-T. *et al*. Adaptive EEG-based alertness estimation system by using ICA-based fuzzy neural networks. *IEEE Transactions on Circuits and Systems I: Regular Papers* **53**, 2469–2476, https://doi.org/10.1109/tcsi.2006.884408 (2006).
54. Penaranda, B. N. & Baldwin, C. L. Temporal factors of EEG and artificial neural network classifiers of mental workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* **56**, 188–192, https://doi.org/10.1177/1071181312561016 (2012).
55. Wilson, G. F. & Russell, C. A. Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Human Factors* **45**, 635–643 Last accessed on 2014-02-18 (2003).
56. Gardony, A. L., Eddy, M. D., Brunyé, T. T. & Taylor, H. A. Cognitive strategies in the mental rotation task revealed by EEG spectral power. *Brain and Cognition* **118**, 1–18, https://doi.org/10.1016/j.bandc.2017.07.003 (2017).
57. Puma, S., Matton, N., Paubel, P.-V., Raufaste, É. & El-Yagoubi, R. Using theta and alpha band power to assess cognitive workload in multitasking environments. *International Journal of Psychophysiology* **123**, 111–120, https://doi.org/10.1016/j.ijpsycho.2017.10.004 (2018).
58. Radüntz, T., Scouten, J., Hochmuth, O. & Meffert, B. Automated EEG artifact elimination by applying machine learning algorithms to ICA-based features. *Journal of Neural Engineering* **14**, 046004, https://doi.org/10.1088/1741-2552/aa69d1 (2017).
59. Mognon, A., Jovicich, J., Bruzzone, L. & Buiatti, M. ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology* **48**, 229–240, https://doi.org/10.1111/j.1469-8986.2010.01061.x (2011).
60. Radüntz, T. Dual frequency head maps: A new method for indexing mental workload continuously during execution of cognitive tasks. *Frontiers in Physiology* **8**, 1019, https://doi.org/10.3389/fphys.2017.01019 (2017).
61. Unsworth, N., Heitz, R. P., Schrock, J. C. & Engle, R. W. An automated version of the operation span task. *Behavior Research Methods* **37**, 498–505, https://doi.org/10.3758/BF03192720 Last accessed on 2014-04-02 (2005).
62. Makeig, S., Bell, A. J., Jung, T.-P. & Sejnowski, T. J. Independent component analysis of electroencephalographic data. In Touretzky, D. S., Mozer, M. C. & Hasselmo, M. E. (eds.) *Advances in Neural Information Processing Systems 8*, 145–151 (MIT Press, 1996).
63. Hjorth, B. An on-line transformation of EEG scalp potentials into orthogonal source derivations. *Electroencephalography and Clinical Neurophysiology* **39**, 526–530, https://doi.org/10.1016/0013-4694(75)90056-5 (1975).
64. Amin, H. U., Malik, A. S., Badruddin, N. & Chooi, W.-T. Brain behavior in learning and memory recall process: A high-resolution EEG analysis. In *IFMBE Proceedings*, 683–686, https://doi.org/10.1007/978-3-319-02913-9_174 (Springer International Publishing, 2014).
65. Huang, H. J., Kram, R. & Ahmed, A. A. Reduction of metabolic cost during motor learning of arm reaching dynamics. *Journal of Neuroscience* **32**, 2182–2190, https://doi.org/10.1523/JNEUROSCI.4003-11.2012 (2012).

Acknowledgements

I would like to thank my student assistants Lea Rabe, Emilia Cheladze, and Friederice Schröder for conducting the experiments and the participants for their contribution during the experiments. Furthermore, I want to thank Marion Freyer for providing computational support for the data analysis with SPSS and graphic editing. Finally, I would also like to thank Beate Meffert for critical reading of the manuscript.

Author contributions

T.R. initiated the project and was responsible for the overall conception of the investigation. T.R. was responsible for the implementation of the tasks and the overall technical support of the study. T.R. was responsible for the signal processing, data analysis and evaluation as well as interpretation of the results. The manuscript was written by T.R.

Competing interests

The author declares that the research was conducted in the absence of any non-financial competing interests and in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Additional information

Correspondence and requests for materials should be addressed to T.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.